Spring 5-15-2015

Nonparametric Bayesian Quantile Regression via Dirichlet Process Mixture Models

Chao Chang
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Mathematics

Dissertation Examination Committee:

Nan Lin, Chair

Siddhartha Chib

Jimin Ding

Todd Kuffner

Mladen Victor Wickerhauser

Nonparametric Bayesian Quantile Regression via Dirichlet Process Mixture Models

by

Chao Chang

A dissertation presented to the
Graduate School of Arts & Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2015

St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor Prof. Nan Lin for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, and immense knowledge. Without his guidance, the completion of this thesis would not have been possible. I sincerely thank Prof. Mladen Victor Wickerhauser for his kindness and willingness to provide help at any time. I would like to thank Prof. Jimin Ding for organizing statistics seminars and colloquium talks, which are very beneficial. I would like to thank Prof. Todd Kuffner for organizing the Probability and Statistics Reading Group. I am indebted to Prof. Siddhartha Chib for his insightful suggestions on my research. I would like to thank Prof. Antonio Galvao at University of Iowa, who kindly offers me the opportunity for collaboration on the quantile regression for longitudinal data.

I would also like to thank the Department of Mathematics for the generous support, without which I could not obtain my Ph.D degree. I am full of gratitude to all faculty in the department, from whom I learned so much beautiful mathematics. In particular, I would like to thank Prof. Stanley Sawyer for his excellent instruction in real analysis. I would like to thank my fellow graduate students for their support and friendship over the years.

Last but not least, I would like to thank my family, especially my parents and my cousin, for their unconditional love and support.

Dedicated to My family.

ABSTRACT OF THE DISSERTATION

Nonparametric Bayesian Quantile Regression via Dirichlet Process Mixture Models

by

Chao Chang

Doctor of Philosophy in Mathematics,

Washington University in St. Louis, 2015.

Professor Nan Lin, Chair

We propose new nonparametric Bayesian approaches to quantile regression using Dirichlet process mixture (DPM) models. All the existing quantile regression methods based on DPMs require the kernel density to satisfy the quantile constraint, hence the kernel densities are themselves usually in the form of mixtures. One innovation of our approaches is that we impose no constraint on the kernel, thus a wide range of densities can be chosen as the kernels of the DPM model. The quantile constraint is satisfied by a post-processing of the DPM by a suitable location shift. As a result, our proposed models use simpler kernels and yet possess great flexibility by mixing over both the location parameter and the scale parameter. The posterior consistency of our proposed model is studied carefully. And Markov chain Monte Carlo algorithms are provided for posterior inference. The performance of our approaches is evaluated using simulated data and real data. Moreover, we are able to incorporate random effects into our models such that our approaches can be extended to handle longitudinal data.

1. Introduction

In this chapter, we introduce the background of quantile regression and Dirichlet process mixture models and motivate the idea of Bayesian quantile regression using Dirichlet process mixture models where our main contribution lies.

1.1 Quantile regression

Mean regression, e.g. linear regression, has been widely used to model the relationship between the covariates and response in a variety of applications. As a powerful complement to the mean regression, quantile regression, proposed in [62], can provide a more complete description of the functional dependence of the response on the covariates. There have been a large volume of literature on the application of quantile regression in various fields like social sciences and econometrics [1, 2, 6, 54, 61, 63, 78].

Let p denote the quantile of interest. Given a random variable V with distribution function $F_V(v)$, define the quantile function as

$$Q_V(p) = \inf\{v : p \leq F_V(v)\}. \tag{1.1}$$

Given the response variable y and the covariate $\boldsymbol{x} \in \mathbb{R}^m$, the quantile regression model can be formulated as $Q_{Y|\boldsymbol{x}}(p) = \boldsymbol{x}^T\boldsymbol{\beta}$, or equivalently, $Y = \boldsymbol{x}^T\boldsymbol{\beta} + \epsilon$ with the quantile constraint $Q_\epsilon(p) = 0$ on the error distribution. Given the data $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$, [62] proposed

a frequentist solution by minimizing the check loss function $\rho_p(u) = u(p - \mathbb{I}(u < 0))$, where $\mathbb{I}$ denotes the indicator function. In other words,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^{\boldsymbol{m}}} \sum_{i=1}^{n} \rho_p(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}). \tag{1.2}$$

One major advantage of quantile regression over mean regression is that it has minimal assumptions on the error distributions except the quantile constraint. While computation for (1.2) is fast as discussed in [64] and Chapter 6 of [61], exact inference can be more difficult to obtain. As a result, frequentist inference for quantile regression is mainly based on either asymptotic theories or bootstrap methods. A comprehensive review can be found in [61] and references therein.

Bayesian methods naturally provide exact inference and have been recently studied extensively for quantile regression. The most straightforward Bayesian quantile regression method was proposed in [118]. This parametric approach assumes that the error distribution follows the asymmetric Laplace distribution (ALD) whose density is given by

$$f(z|\mu, \sigma, p) = \frac{p(1-p)}{\sigma} \exp\left\{-\rho_p\left(\frac{z-\mu}{\sigma}\right)\right\}.$$

In this way, the maximum a posteriori estimate agrees with the estimate in (1.2). A generalization to incorporate regularization was proposed in [72]. This parametric Bayesian quantile regression approach has been widely used, for example [36, 71, 80, 101, 119]. However, modelling the error distribution directly as the ALD is very restrictive. The probability density function (pdf) of the ALD is always non-differentiable at one point and is skewed except when median is the quantile of interest.

To avoid the restrictive parametric assumption, considerable efforts have been devoted to applying nonparametric Bayesian approaches to model the error distribution in a more flexible fashion. One popular approach is based on Dirichlet process mixture (DPM)

models, which will be reviewed in the next section. [65] developed median regression models using DPMs. [99] performed quantile regression by jointly modelling the response and the covariate using a DPM of multivariate normal distributions. [66, 89] proposed quantile regression based on DPMs by designing kernel densities that satisfy the quantile constraint.

Besides the DPM-based methods, there are other branches of nonparametric Bayesian techniques for quantile regression. [117] proposed to solve the quantile regression problem using Bayesian empirical likelihood proposed in [70]. This method has computational challenges due to the multi-modality of the empirical likelihood and the empirical likelihood ratio. [68] further proposed an approach to quantile regression based on the Bayesian exponentially tilted empirical likelihood developed in [92]. And [26, 27] incorporated the idea in [69] to propose approximate methods for quantiles based on substitution likelihood originally developed in [55]. Inference in [26, 27, 68] relies on the asymptotic normality of the posterior, and thus can be unreliable for moderate or small data. A method based on approximating the likelihood by a linear interpolation of the quantiles was proposed in [30]. One downside of this method is the slow convergence of the proposed Metropolis-Hastings (MH) algorithm. All these methods have computational issues due to the lack of nicely designed Markov chain Monte Carlo (MCMC) algorithms.

1.2 Dirichlet process mixture models

First we define the Dirichlet process originally introduced by Ferguson in [31,32]. Let $\mathcal{X}$ be a standard Borel space with Borel σ-algebra $\mathcal{A}$ and $\mathcal{P}$ be the space of probability measures on $(\mathcal{X}, \mathcal{A})$ equipped with the weak topology and the corresponding Borel σ-algebra $\mathcal{M}$.

Definition 1.2.1 *A random measure P on $(\mathcal{X}, \mathcal{A})$ is said to have a Dirichlet process distribution $DP(\alpha, G)$ with concentration parameter α and base measure G, if for every finite measurable partition $A_1, \cdots, A_k$ of $\mathcal{X}$,*

$$(P(A_1), \cdots, P(A_k)) \sim Dirichlet(\alpha G(A_1), \cdots, \alpha G(A_k)),$$

where $G(\cdot)$ is a probability measure on $\mathcal{X}$.

The Dirichlet process defines a probability measure on $(\mathcal{P}, \mathcal{M})$ and has some nice properties. Given $P \sim DP(\alpha, G)$, for any $A \in \mathcal{A}$,

$$E(P(A)) = G(A) \text{ and } Var(P(A)) = \frac{G(A)(1 - G(A))}{1 + \alpha}.$$

That is, the mean of P is equal to the base measure G and α controls how similar P is to the base measure G. Further, the Dirichlet process is a conjugate prior. If $\theta_1, \cdots, \theta_n \sim P$ with $P \sim DP(\alpha, G)$, then

$$P|\theta_1, \cdots, \theta_n \sim DP\left(\alpha + n, \ \frac{\alpha}{\alpha + n} G + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{\theta_i}\right), \tag{1.3}$$

where δ_θ denotes the Dirac delta measure, that is, for any $A \subseteq \mathcal{X}$,

$$\delta_\theta(A) = \mathbb{I}_A(\theta) = \begin{cases} 1, & \theta \in A; \\ 0, & \theta \notin A. \end{cases}$$

Sethuraman [95] introduced an important method for constructing a Dirichlet process, which is also referred as the stick-breaking representation. Let $\theta_1, \theta_2, \cdots \sim G$ and $V_1, V_2, \cdots \sim \text{Beta}(1, \alpha)$ be mutually independent. Let $p_i = V_i \prod_{j=1}^{i-1}(1 - V_j)$, $i \in \mathbb{N}^+$. Then

$$P = \sum_{i=1}^{\infty} p_i \delta_{\theta_i} \sim DP(\alpha, G). \tag{1.4}$$

In [9], Blackwell and MacQueen presented the Pólya urn scheme to sample from $P \sim DP(\alpha, G)$ by integrating out P,

$$\theta_i|\theta_1, \cdots, \theta_{i-1} \sim \frac{1}{i - 1 + \alpha} \sum_{j=1}^{i-1} \delta_{\theta_j} + \frac{\alpha}{i - 1 + \alpha} G. \tag{1.5}$$

Due to the conjugacy of the Dirichlet process, the Pólya urn scheme essentially provides an algorithm for the posterior inference of the Dirichlet process. Also from (1.5), the samples drawn are naturally clustered, which makes the Dirichlet process a popular method for clustering [100].

Although the Dirichlet process is a popular nonparametric Bayesian method and has broad applications [5,31,32,100], one obvious limitation of the Dirichlet process is that the generated probability measure is almost surely discrete, thus a Dirichlet process cannot be used as a prior for estimating a density. To overcome the discreteness, [33] proposed the DPM of normal distributions and [74] developed a class of priors on densities using the DPM of any known densities. Let $k(z|\boldsymbol{\theta})$ with parameters $\boldsymbol{\theta} \in S \subseteq \mathbb{R}^l$ be a kernel, that is, for each $\boldsymbol{\theta}$, $k(\cdot|\boldsymbol{\theta})$ is a probability density function. Given a Dirichlet process $P \sim DP(\alpha, G)$ with G being a probability measure over S, we can define a probability measure over the space of densities by

$$f(z) = \int k(z|\boldsymbol{\theta}) dP(\boldsymbol{\theta}), \quad P \sim DP(\alpha, G).$$

DPM models are very appealing in Bayesian density estimation and regression [22,29], because they are very flexible and also computationally simple in the sense that there are well developed MCMC sampling methods for DPM models, for example [29, 49, 77, 82, 109, 113]. For DPM models, choice of the kernel distribution usually has a profound impact on the efficiency. In the next section, we will elaborate this issue in the context of quantile regression.

1.3 Bayesian quantile regression via Dirichlet process mixture models

When modelling the error distribution in quantile regression by a DPM, it has to satisfy the quantile constraint. A natural solution is to let the kernel densities satisfy the

quantile constraint, then after an application of Fubini's theorem, the resulted DPM also meets the quantile constraint. [65,66,89] considered kernels including the ALD and some two-component mixture distributions.

In [66], Kottas and Kranjajić considered the scale DPM of the ALD as in (1.6) and the scale DPM of mixtures of two uniform distributions as in (1.7).

$$f_\epsilon(z) = \int \frac{p(1-p)}{\sigma} \exp\left\{-\rho_p\left(\frac{z}{\sigma}\right)\right\} dP(\sigma), \quad P \sim DP(\alpha, G). \tag{1.6}$$

$$\begin{aligned} f_\epsilon(z) &= \int\int \left(\frac{p}{\sigma_1}\mathbb{I}_{(-\sigma_1,0)}(z) + \frac{1-p}{\sigma_2}\mathbb{I}_{[0,\sigma_2)}(z)\right) dP_1(\sigma_1)dP_2(\sigma_2), \\ & P_r \sim DP(\alpha_r, G_r), \quad r = 1, 2. \end{aligned} \tag{1.7}$$

Although the quantile constraint is satisfied, the proposed models in [66] are relatively restrictive as shown in Fig 1.1. Firstly, distributions for (1.6) and (1.7) always have modes at the quantile of interest. Secondly, the mixture of the ALD (1.6) is not differentiable and the mixture of uniform distribution (1.7) is not even continuous. Thirdly, both mixtures can only be unimodal.

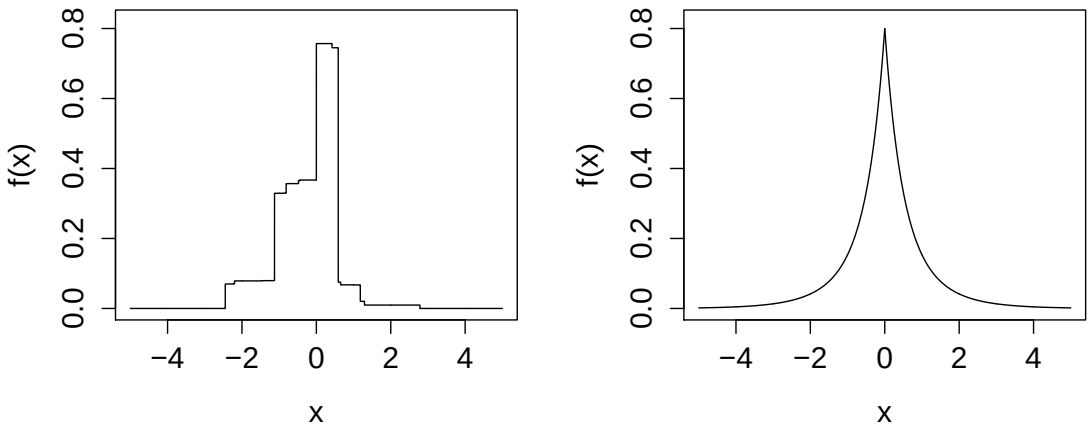
[89] specified the kernel distribution as a mixture of two normal distributions,

$$\begin{aligned} k(z|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) &= q_{\mu_1,\mu_2,\sigma_1^2,\sigma_2^2}\phi(z|\mu_1, \sigma_1^2) + (1 - q_{\mu_1,\mu_2,\sigma_1^2,\sigma_2^2})\phi(z|\mu_2, \sigma_2^2), \\ \text{and} \quad q_{\mu_1,\mu_2,\sigma_1^2,\sigma_2^2} &= \frac{p - \Phi(-\mu_2/\sigma_2)}{\Phi(-\mu_1/\sigma_1) - \Phi(-\mu_2/\sigma_2)}, \end{aligned} \tag{1.8}$$

where $\phi(z|\mu.\sigma^2)$ denotes the pdf of the normal distribution with mean μ and variance σ^2 and Φ denotes the cumulative distribution function of the standard normal distribution. This method is more flexible than (1.6) and (1.7) and is able to model distributions with arbitrary skewness as well as multi-modality. However, we believe this method is overcomplicated in that it requires to mix over four parameters, which may lead to unnecessary computational burden and reduce the efficiency in estimating the regression coefficient. And due to the complexity of the kernel, for ease of implementation, this

DPM model is approximated by a truncated DPM model, i.e. a finite mixture model, which leads to loss of accuracy in estimation. Besides, since the kernel distribution is itself a mixture, the resulting cluster information, as a standard by-product of Dirichlet processes [23, 100], is hard to interpret.

Figure 1.1. Error densities estimated using DPM of mixtures of uniform distribution (left) and DPM of ALD (right) for $p = 0.5$.



One important innovation of this thesis is to use kernels that do not satisfy the quantile constraint. Even if the kernel density violate the quantile constraint, the DPM may not. For example, let the kernel $k(z|\theta)$ be a location family, that is, $k(z|\theta) = k(z - c|\theta - c)$ for any $c \in \mathbb{R}$. Given a mixing distribution P, find $q \in \mathbb{R}$ such that $\int_{-\infty}^{q} \int k(z|\theta) dP(\theta) dz = p$. Now we can define another kernel $k_0(z|\theta) = k(z|\theta - q)$. Then by Fubini's theorem and a change of variable, we can see that the mixture $\int k_0(x|\theta) dP(\theta)$ satisfies the quantile constraint, while the kernel $k_0(x|\theta)$ does not almost surely.

Our contribution is threefold. First, we propose a novel DPM-based method with the kernel density being a single location-scale logistic distribution. While enjoying the simplicity on the kernel, our mixture model also provides great flexibility by mixing over

both the location parameter and the scale parameter. Without any adjustment, there is no guarantee that the quantile constraint is satisfied. However, we carefully study how the constraint impacts the inference of the regression parameters and we are able to provide a simple adjustment to get correct inference of the regression coefficients even when the quantile constraint is violated. And we are able to show that our proposed model is equivalent to the model which employs location shift of the mixture to satisfy the quantile constraint. We thus avoid the complication of using a mixture kernel density to satisfy the quantile constraint.

Secondly, we establish the theoretical guarantee of the posterior consistency on the regression coefficients and density estimation for our proposed models. Here, we define posterior consistency in the context of nonparametric Bayesian regression. For simplicity, assume the regression model is $Y = \beta_0^* + \beta_1^* x + \epsilon$ with the true error density f^*. Let $\mathcal{F}$ denote the space of densities. For any pdf f, let P_f denote the corresponding probability measure.

Definition 1.3.1 *Given an infinite sequence of fixed covariates* $\{x_i\}_{i=1}^{\infty}$ *and a prior* Π *over* $\mathcal{F} \times \mathbb{R} \times \mathbb{R}$. *For each* n, *let* $\Pi(\cdot|\mathbf{Y}_n)$ *denote a posterior distribution given data* $Y_1, \cdots, Y_n$. *And let* $f_i^*(y) = f^*(y - \beta_0^* - \beta_1^* x_i)$. *The sequence* $\{\Pi(\cdot|\mathbf{Y}_n)\}$ *is said to be weakly consistent at* $(f^*, \beta_0^*, \beta_1^*)$ *if for any* $\eta > 0$ *and any weak neighbourhood* $\mathcal{U}$ *of* f^*, *as* $n \to \infty$,

$$\Pi\{(f, \beta_0, \beta_1) : f \in \mathcal{U}, |\beta_0 - \beta_0^*| < \eta, |\beta_1 - \beta_1^*| < \eta | \mathbf{Y}_n\} \to 1 \tag{1.9}$$

almost surely $\prod_{i=1}^{\infty} P_{f_i^*}$.

Currently, there is no theory for the posterior consistency in nonparametric Bayesian quantile regression. Since our method involves a location shift depending on the mixing

distribution, substantial modification is required on the tools provided in [4, 102] for proving posterior consistency.

Thirdly, extending the idea of satisfying quantile constraint by a proper location shift, we propose using the DPM of normal distributions (DPMN) for quantile regression, which is computationally more efficient because of the full conjugacy. And we develop a quantile regression model using DPMN for longitudinal data.

This thesis is organized as follows. Chapter 2 introduces the DPM of logistic distribution (DPML) for quantile regression. Chapter 3 develops the posterior consistency theory for DPML. Chapter 4 extends the DPML model to handle data with heteroscedasticity. Chapter 5 presents the DPMN model for longitudinal data. Finally we conclude in Chapter 6.

2. Bayesian quantile regression via DPM of logistic distributions

Given the covariates $\boldsymbol{x}_i$, a $(m+1)$-vector, and the response y_i for $i = 1, \cdots, n$, the linear quantile regression model is formulated as $Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i$ with $Q_\epsilon(p) = 0$, where p is the quantile of interest and $Q_V(\cdot)$ denotes the quantile function of the random variable V as defined in (1.1). When there is no confusion, we omit the subscript V and let $Q(\cdot)$ denote the quantile function. Following convention, we let β_0 be the intercept, so correspondingly $x_{i0} = 1$ for $i = 1, \cdots, n$. In the sequel, we always set the quantile of interest as the p-th quantile.

For most commonly used distributions, the quantile function $Q(\cdot)$ does not have a closed form and is expressed as the solution to the equation in the form of an integral $\int_{-\infty}^{Q(p)} f(x)dx = p$, where f is the corresponding pdf. Among a few exceptions, the logistic distribution, whose pdf is $f(x) = \frac{\exp\left(-\frac{x-\mu}{\sigma}\right)}{\sigma\left(1+\exp\left(-\frac{x-\mu}{\sigma}\right)\right)^2}$, enjoys a simple quantile function, $Q(p) = \mu + \sigma \log \frac{p}{1-p}$, where μ and σ are the mean and scale parameter, respectively. Thus we can explicitly re-parametrize the logistic distribution by its p-th quantile τ and the scale parameter σ. Denote the logistic density with $\tau = 0$ and $\sigma = 1$ by $\psi(x) := \frac{\frac{1-p}{p}\exp(-x)}{\left(1+\frac{1-p}{p}\exp(-x)\right)^2}$. And if a random variable V has the pdf $\frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right)$, we denote by $V \sim \text{Logistic}(\tau, \sigma)$.

Using some common choices on the prior, we summarize our proposed model as follows.

$$
\begin{aligned}
& y_i|\tau_i, \sigma_i, \boldsymbol{\beta}, \boldsymbol{x}_i \sim \text{Logistic}(\boldsymbol{x}_i^T\boldsymbol{\beta} - \tau_i, \sigma_i), \quad i = 1, \cdots, n, \\
& \tau_i, \sigma_i|P \sim P, \quad i = 1, \cdots, n, \\
& P|\alpha, G \sim DP(\alpha, G), \\
& G(\tau, \sigma) = \text{Logistic}(\tau| - \sigma \log \lambda, \sigma) \cdot \text{Inv-Gamma}(\sigma|c, d), \\
& \beta_i \overset{i.i.d.}{\sim} \text{N}(0, \nu), \quad i = 0, \cdots, m, \\
& \alpha \sim \text{Gamma}(a_1, b_1), \\
& d \sim \text{Gamma}(a_2, b_2),
\end{aligned}
\tag{2.1}
$$

where c, ν, a_1, a_2, b_1 and b_2 are hyper-parameters and λ is the solution to Equation (2.3).

Note that our kernel density Logistic$(-\tau_i, \sigma_i)$ has its p-th quantile equal to $-\tau_i$, which may not be 0, so the quantile constraint is violated. But as discussed in the next two sections, our proposed model is still valid for the inference on the regression coefficients. This a novel specification, as in the literature [65, 66, 89], the kernel densities are always required to satisfy the quantile constraint.

2.1 DPM of logistic distributions

Given a kernel density $k(z|\boldsymbol{\theta})$ with parameters $\boldsymbol{\theta} \in S \subseteq \mathbb{R}^l$ and a Dirichlet process $P \sim DP(\alpha, G)$ with G being a probability measure over S, DPM [33, 74] defines a probability measure over the space of pdfs by $g(z) = \int k(z|\boldsymbol{\theta})dP(\boldsymbol{\theta})$ with $P \sim DP(\alpha, G)$.

In the context of quantile regression, if the pdf for the error ϵ_i's, g_ϵ, is modelled by a DPM with kernel density k, we want to choose a kernel k such that $\int_{-\infty}^{0} g_\epsilon(z)dz = p$. A simple solution is to choose k which satisfies the quantile constraint itself. Then by Fubini's theorem, g_ϵ is also guaranteed to satisfy the quantile constraint. Existing DPM-

based quantile regression methods [65,66,89] are all based on this simple idea, which leads to either inflexible models as in (1.6) and (1.7) or overly complex models as in (1.8).

We argue that the quantile constraint requirement on the kernel densities is not necessary. We can use a simpler kernel that violates the quantile constraint, as long as it is guaranteed that the resulting mixture satisfies the quantile constraint.

The pdf for the error ϵ_i's can be modelled by the DPM of the logistic densities,

$$f_\epsilon(z) = \int \frac{1}{\sigma}\psi\left(\frac{z+\tau}{\sigma}\right) dP(\tau,\sigma), \quad P \sim DP(\alpha, G). \tag{2.2}$$

Our first attempt is to specify an appropriate base measure G, such that by integrating out the Dirichlet process, the mean error density satisfies the quantile constraint. We introduce a property of the logistic distribution.

Proposition 2.1.1 *If two random variables X and Y are independent, $X \sim$ Logistic$(0,\sigma)$ and $Y \sim$ Logistic$(-\sigma\log\lambda,\sigma)$, with λ being the solution to the equation*

$$x\log x + 1 - x - p(1-x)^2 = 0 \tag{2.3}$$

subject to $\lambda = 1$ if $p = 0.5$ and $\lambda \neq 1$ if $p \neq 0.5$, then

(1) $Q_{X-Y}(p) = 0$,

(2) such λ exists for any $p \in (0,1)$ and is unique.

Proof (1) Assume $Y \sim Logistic(\delta, \sigma)$, by requiring $Q_{X-Y}(p) = 0$, we have

$$
\begin{aligned}
p = P(X - Y \leq 0) &= \int_{-\infty}^{\infty} \int_{-\infty}^{y} f_{X,Y}(x, y) dx dy \\
\text{(By independence)} \ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{y} f_X(x) dx \right) f_Y(y) dy \\
&= \int_{-\infty}^{\infty} \frac{1}{1 + \frac{1-p}{p} \exp\left(-\frac{y}{\sigma}\right)} \frac{\frac{1-p}{p} \exp\left(-\frac{y-\delta}{\sigma}\right)}{\sigma\left(1 + \frac{1-p}{p} \exp\left(-\frac{y-\delta}{\sigma}\right)\right)^2} dy \\
&\left[\text{Set } t = \frac{1-p}{p} \exp\left(-\frac{y-\delta}{\sigma}\right), \ \lambda = \exp\left(-\frac{\delta}{\sigma}\right) \right] \\
&= \int_{0}^{\infty} \frac{1}{(1 + \lambda t)(1 + t)^2} dt \\
&= \begin{cases} 0.5 & \text{if } \lambda = 1, \\ \frac{\lambda \log \lambda + 1 - \lambda}{(\lambda - 1)^2} & \text{otherwise.} \end{cases}
\end{aligned}
$$

So to guarantee the quantile constraint, $\delta = -\sigma \log \lambda$ where λ satisfies $\lambda \log \lambda + 1 - \lambda - p(1 - \lambda)^2 = 0$ subject to $\lambda = 1$ when $p = 0.5$ and $\lambda \neq 1$ otherwise.

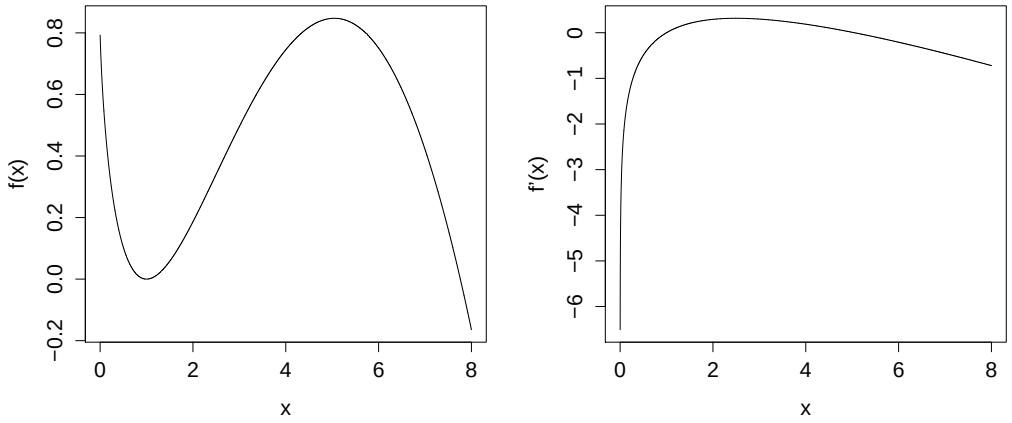
(2) Now we show the existence and uniqueness of such λ. If $p = 0.5$, clearly λ exists and is unique. If $p \neq 0.5$, it suffices to show the equation $x \log x + 1 - x - p(1 - x)^2 = 0$ has only one solution besides 1. Define a function $f(x) = x \log x + 1 - x - p(1 - x)^2$. Then $f'(x) = \log x - 2p(x - 1)$.

We first show that $f'(x)$ has two roots. Since $f''(x) = \frac{1}{\lambda} - 2p$ and $f'''(x) = -1/x^2 < 0$, f' achieves its maximum at $x = \frac{1}{2p}$, and its maximum value is $2p - 1 - \log(2p)$. By considering the function $g(x) = x - 1 - \log x$, it is easy to show that $2p - 1 - \log(2p) > 0$ when $p \neq 0.5$. Observe that $f'(0) = -\infty$ and $f'(\infty) = -\infty$. Therefore, $f'(\frac{1}{2p}) > 0$ implies that $f'(x)$ has at least two roots. And consider the graphs of $y = \log x$ and $y = 2p(x - 1)$, they can have at most two intersections. Then $f'(x)$ can have at most two roots. Thus,

$f'(x)$ has exactly two roots and we know one of them is equal to 1. Denote the other root of $f'(x)$ by x_0.

Now if $p < 0.5$, then $f'(x)$ achieves its maximum at $\frac{1}{2p} > 1$, which implies $x_0 > 1$. So $f'(x) < 0$ on $(0, 1)$, $f'(x) > 0$ on $(1, x_0)$ and $f'(x) < 0$ on (x_0, ∞). That is, $f(x)$ is decreasing on $(0, 1)$, increasing on $(1, x_0)$ and decreasing on (x_0, ∞). We know $f(0) = 1 - p > 0$, $f(1) = 0$ and $f(\infty) = -\infty < 0$, so $f(x)$ has another root in (x_0, ∞). Figure 2.1 plots $f(x)$ and $f'(x)$ for $p = 0.2$.

Figure 2.1. The graph of f (left) and the graph of f' (right) for $p = 0.2$.



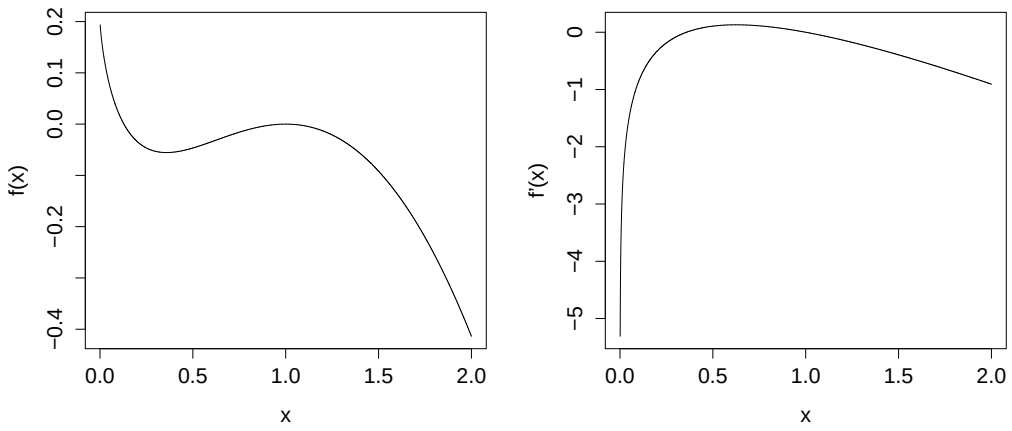
If $p > 0.5$, then $f'(x)$ achieves its maximum at $\frac{1}{2p} < 1$, which implies $x_0 < 1$. So $f'(x) < 0$ on $(0, x_0)$, $f'(x) > 0$ on $(x_0, 1)$ and $f'(x) < 0$ on $(1, \infty)$. That is, $f(x)$ is decreasing on $(0, x_0)$, increasing on $(x_0, 1)$ and decreasing on $(1, \infty)$. Again since we know $f(0) = 1 - p > 0$, $f(1) = 0$ and $f(\infty) = -\infty < 0$, $f(x)$ has another root in $(0, x_0)$. Figure 2.2 plots $f(x)$ and $f'(x)$ for $p = 0.8$.

So we conclude that $f(x)$ has exactly two roots, then the proof completes. ■

Now we set the base measure as

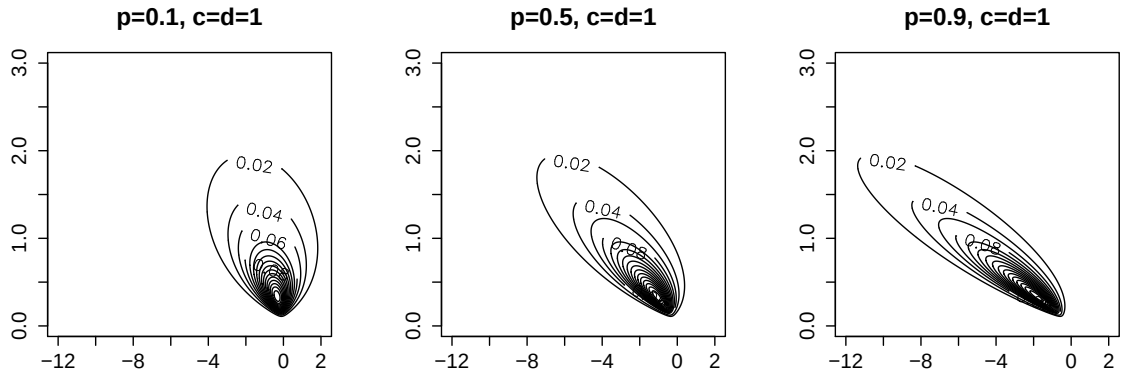
$$G(\tau, \sigma) = \text{Logistic}(\tau | -\sigma \log \lambda, \sigma) \cdot \text{Inv-Gamma}(\sigma | c, d), \tag{2.4}$$

Figure 2.2. The graph of f (left) and the graph of f' (right) for $p = 0.8$.



where λ is given in Proposition 2.1.1 and c and d are hyperparameters. Figure 2.3 plots the contour plots of G. This base measure G has the following desired property.

Figure 2.3. Contour plot for the base measure G.



Proposition 2.1.2 *Let $\tilde{\Pi}$ denote $DP(\alpha, G)$ with G defined in (2.4), then we have*

$$\int_{-\infty}^{0} \int \int \frac{1}{\sigma} \psi \left(\frac{z+\tau}{\sigma} \right) dP(\tau, \sigma) d\tilde{\Pi}(P) dz = p.$$

Proof If $P \sim DP(\alpha, G)$, by Definition 1.2.1, for any measurable set $A \subseteq \mathbb{R} \times \mathbb{R}^+$, $P(A) \sim \text{Beta}(\alpha G(A), \alpha[1 - G(A)])$. Thus $E(P(A)) = G(A)$, that is, for any measurable set $A \subseteq \mathbb{R} \times \mathbb{R}^+$,

$$\int \int \mathbb{I}_A(\tau, \sigma) dP(\tau, \sigma) d\tilde{\Pi} = \int \mathbb{I}_A(\tau, \sigma) dG(\tau, \sigma),$$

where $\mathbb{I}$ denotes the indicator function. Then for any simple function $g(\tau, \sigma)$, which is a finite linear combinations of indicator functions of measurable sets,

$$\int \int g(\tau, \sigma) dP(\tau, \sigma) d\tilde{\Pi} = \int g(\tau, \sigma) dG(\tau, \sigma).$$

Therefore,

$$\int \int \frac{1}{\sigma} \psi \left(\frac{z + \tau}{\sigma} \right) dP(\tau, \sigma) d\tilde{\Pi} = \int \frac{1}{\sigma} \psi \left(\frac{z + \tau}{\sigma} \right) dG(\tau, \sigma).$$

Now let h denote the density of the Inv-Gamma$(2, d)$ distribution. By the definition of G in (2.1), it follows from Fubini's theorem that

$$\begin{aligned}
&\int_{-\infty}^{0} \int \int \frac{1}{\sigma} \psi \left(\frac{z + \tau}{\sigma} \right) dP(\tau, \sigma) d\tilde{\Pi}(P) dz \\
&= \int_{-\infty}^{0} \int \frac{1}{\sigma} \psi \left(\frac{z + \tau}{\sigma} \right) dG(\tau, \sigma) dz \\
&= \int_{-\infty}^{0} \int_{0}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\sigma} \psi \left(\frac{z + \tau}{\sigma} \right) \frac{1}{\sigma} \psi \left(-\frac{z + \sigma \log \lambda}{\sigma} \right) h(\sigma) d\tau d\sigma dz \\
&= \int_{0}^{\infty} \left(\int_{-\infty}^{0} \int_{-\infty}^{\infty} \frac{1}{\sigma} \psi \left(\frac{z + \tau}{\sigma} \right) \frac{1}{\sigma} \psi \left(-\frac{z + \sigma \log \lambda}{\sigma} \right) d\tau dz \right) h(\sigma) d\sigma \\
&= \int_{0}^{\infty} p h(\sigma) d\sigma = p.
\end{aligned}$$

The fourth equality follows from Proposition 2.1.1. ■

Proposition 2.1.2 shows that the quantile constraint is satisfied for the prior mean error density after integrating out the Dirichlet process, although conditioning on the Dirichlet

process, the quantile constraint is almost surely violated. As Bayesian estimates are often based on the posterior mean, to get correct inference, it is sufficient to have the posterior mean error density satisfy the quantile constraint. However, this is not the case. Fortunately, even when the quantile constraint is violated, we can still get correct inference for the regression coefficients after some simple adjustment proposed in the next subsection, where we carefully study how the constraint on the location parameter relates to the posterior inference for the regression coefficients.

2.2 Adjustment

In this subsection we will show that violating the quantile constraint does not affect the estimation of the regression coefficients except for the intercept. And a simple adjustment can be made to correct the estimation for the intercept.

To make the above argument precise, we first introduce some notations and discuss the problem in a more general setting. For simplicity, consider a regression model with a univariate covariate, $Y = \beta_0 + \beta_1 x + \epsilon$ satisfying the quantile constraint $Q_\epsilon(p) = 0$. The observations are (y_i, x_i), $i = 1, \cdots, n$. We let $\pi_1(\beta_0)$ and $\pi_2(\beta_1)$ denote the independent priors for the regression coefficients. Also we assume the support of π_1 and π_2 are both $(-\infty, \infty)$. Let Λ denote any probability measure over the space of probability measures that are absolutely continuous with respect to the Lebesgue measure. We have the first model as

$$(A)\ \ y_i - \beta_0 - \beta_1 x_i | \beta_0, \beta_1, x_i \sim F \text{ with the pdf } f_F \text{ for } i = 1, \cdots, n,$$

$$\beta_0 \sim \pi_1, \ \ \beta_1 \sim \pi_2 \text{ and } F \sim \Lambda.$$

Obviously, the quantile constraint may be violated in model (A). For each f_F, define q_F such that $f_F(z - q_F)$ satisfies the quantile constraint. The existence of such q_F is directly

from the fact that the quantile constraint is on the location parameter of f_F. Then we can define a random probability measure Λ^* based on Λ. We say $F^* \sim \Lambda^*$ if and only if there exists $F \sim \Lambda$ such that the pdf of F^* is given by $f_{F^*}(z) = f_F(z - q_F)$. So we have another model

$$(B)\ \ y_i - \beta_0 - \beta_1 x_i | \beta_0, \beta_1, x_i \sim F^* \text{ with the pdf } f_{F^*} \text{ for } i = 1, \cdots, n,$$

$$\beta_0 \sim \pi_1, \ \ \beta_1 \sim \pi_2 \text{ and } F^* \sim \Lambda^*.$$

By definition, each f_{F^*} in model (B) satisfies the quantile constraint.

Let $E^{(A)}(\beta_0|\mathbf{x}, \mathbf{y})$ and $E^{(B)}(\beta_0|\mathbf{x}, \mathbf{y})$ denote the posterior mean of β_0 in models (A) and (B), respectively. Let $E^{(A)}(q_F|\mathbf{x}, \mathbf{y})$ denote the posterior mean of q_F in model (A). Also let $Var^{(B)}(\beta_0|\mathbf{x}, \mathbf{y})$ denote the posterior variance of β_0 in model (B) and let $Var^{(A)}(\beta_0 - q_F|\mathbf{x}, \mathbf{y})$ denote the posterior variance of $\beta_0 - q_F$ in model (A).

Proposition 2.2.1 *If the prior of β_0 is $\pi_1(\beta_0) \propto 1$,*

(1) the posterior distribution of β_1 in (A) is the same as that in (B);

(2) $E^{(B)}(\beta_0|\mathbf{x}, \mathbf{y}) = E^{(A)}(\beta_0|\mathbf{x}, \mathbf{y}) - E^{(A)}(q_F|\mathbf{x}, \mathbf{y})$;

(3) $Var^{(B)}(\beta_0|\mathbf{x}, \mathbf{y}) = Var^{(A)}(\beta_0 - q_F|\mathbf{x}, \mathbf{y})$.

These results follow by repeatedly applying the change of variables technique and using the condition $\pi_1(\beta_0) \propto 1$. The detail of the proof is given in Appendix A.1.

Proposition 2.2.1 can be easily extended to the case with more covariates. Note that in practice, the improper prior for β_0 may be replaced by a normal distribution with a large variance ν, and the estimation error for regression coefficients using the posterior mean is bounded by C/ν with C being a constant not depending on ν. So practically our model (2.2) can be treated as an example of model (A).

Suppose that we have MCMC samples from model (A) for the required location shift q_F and regression coefficients $(\beta_0^{(1)}, \beta_1^{(1)}, q_F^{(1)}), \cdots, (\beta_0^{(T)}, \beta_1^{(T)}, q_F^{(T)})$. According to Proposition 2.2.1, to get correct inference for β_0 and β_1, we should work with the adjusted sample $(\beta_0^{(1)} - q_F^{(1)}, \beta_1^{(1)}), \cdots, (\beta_0^{(T)} - q_F^{(T)}, \beta_1^{(T)})$. For example, we should use $\sum_{t=1}^{T}(\beta_0^{(t)} - q_F^{(t)})/T$ to estimate β_0, and the credible interval should be constructed using $\beta_0^{(1)} - q_F^{(1)}, \cdots, \beta_0^{(T)} - q_F^{(T)}$. Thus after the aforementioned adjustment, the prior (2.2) still provides correct inference for the regression parameter, even though the quantile constraint is violated. With the adjustment, our proposed model is equivalent to model the error density as follows,

$$f_\epsilon(z) = \int \frac{1}{\sigma}\psi\left(\frac{z+\tau-q_P}{\sigma}\right) dP(\tau,\sigma), \quad P \sim DP(\alpha, G), \tag{2.5}$$

where q_P is defined in such a way that $\int_{-\infty}^{q_P} \int \frac{1}{\sigma}\psi\left(\frac{z+\tau-q_P}{\sigma}\right) dP(\tau,\sigma)dz = p$. That is, our model can be thought of as a DPM model which employs a location shift of the mixture to satisfy the quantile constraint in the same fashion as in model (B). We also remark that existing DPM-based methods can be viewed as special cases of model (B) with $q_F \equiv 0$.

2.3 MCMC posterior inference

In this section, we detail a Gibbs sampling scheme for the posterior inference of model (2.1). For each iteration of the Markov chain, we need to update (i) the precision parameter α, (ii) the scale parameter d in the base measure, (iii) the regression coefficients, and (iv) the pairs of location and scale parameters for each sample $\{(\tau_i, \sigma_i)\}_{i=1}^n$ where n is the sample size. Let n^* denote the number of clusters which is equal to the number of distinct pairs in $\{(\tau_i, \sigma_i)\}_{i=1}^n$. And let $\{(\tau_j^*, \sigma_j^*)\}_{j=1}^{n^*}$ denote the distinct pairs.

(i) The full conditional distribution for α is hard to get directly, but as a standard trick [29], one can introduce a fictitious parameter η with prior $U(0,1)$ and update α together with η. Let $\pi_{\eta,N^*} = \frac{a_1+N^*-1}{a_1+N^*-1+N(b_1-\log(\eta))}$, then

$$\alpha|\eta, N^* \sim \begin{cases} \text{Gamma}(a_1+N^*, b_1-\log(\eta)), & \text{with probability } \pi_{\eta,N^*}; \\ \text{Gamma}(a_1+N^*-1, b_1-\log(\eta)), & \text{with probability } 1-\pi_{\eta,N^*}; \end{cases}$$

and

$$\eta|\alpha, N^* \sim \text{Beta}(\alpha+1, N). \tag{2.6}$$

(ii) The full conditional distribution for d is given by

$$d|N^*, \sigma_1^{2*}, \cdots, \sigma_{N^*}^{2*} \sim \text{Gamma}\left(a_2+2N^*, b_2+\sum_{j=1}^{N^*}\frac{1}{\sigma_j^{2*}}\right). \tag{2.7}$$

(iii) The full conditional distributions for the regression coefficients are complicated. Let $u_i = y_i - \boldsymbol{x}_i^T\boldsymbol{\beta} + x_{ik}\beta_k$. Then

$$f(\beta_k|\beta_1,\cdots,\hat{\beta}_k,\cdots,\beta_m,\boldsymbol{\sigma},\boldsymbol{\tau},\boldsymbol{y}) \propto \frac{\exp\left(-\frac{\beta_k^2}{2\nu}-\sum_{i=1}^{n}\frac{u_i-x_{ik}\beta_k+\tau_i}{\sigma_i}\right)}{\prod_{i=1}^{n}\left(1+\frac{1-p}{p}\exp\left(-\frac{u_i-x_{ik}\beta_k+\tau_i}{\sigma_i}\right)\right)^2}.$$

It is easy to verify that the density is log-concave. Here a function $f:\mathbb{R}\to\mathbb{R}^+$ is called log-concave if $f(\xi x+(1-\xi)y) \geq f(x)^\xi f(y)^{1-\xi}$ for all x, y in the domain of f and $0<\xi<1$. So we can apply the adaptive rejection sampling proposed in [45, 46].

(iv) We split this step into two parts. First we update unique pairs (τ_j^*, σ_j^*), and then we update the cluster configuration. The full conditional distributions for (τ_j^*, σ_j^*)'s are complicated. Let $e_i = y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}$, $A = \{i : \tau_i = \tau_j^*, \sigma_i = \sigma_j^*\}$. Also let $|A|$ denote the size of set A and let $A(k)$ denote the kth element in A. Then,

$$f(\tau_j^*|\sigma_j^*, \boldsymbol{\beta}, \boldsymbol{y}) \propto \frac{\exp\left(-\frac{1}{\sigma_j^*}(|A|+1)\tau_j^*\right)}{\prod_{k=1}^{|A|}\left(1+\frac{1-p}{p}\exp\left(-\frac{e_{A(k)}+\tau_j^*}{\sigma_j^*}\right)\right)^2\left(1+\frac{1-p}{p}\exp\left(-\frac{\tau_j^*+\sigma_j^*\log\hat{\lambda}}{\sigma_j^*}\right)\right)^2},$$

and

$$f(\sigma_j^*|\tau_j^*, \boldsymbol{\beta}, \boldsymbol{y}) \propto \frac{\exp\left(-\frac{1}{\sigma_j^*}((|A|+1)\tau_j^* + \sum_{k=1}^{|A|} e_{A(k)} + d)\right)(\sigma_j^*)^{-(|A|+4)}}{\prod_{k=1}^{|A|}\left(1+\frac{1-p}{p}\exp\left(-\frac{e_{A(k)}+\tau_j^*}{\sigma_j^*}\right)\right)^2\left(1+\frac{1-p}{p}\exp\left(-\frac{\tau_j^*+\sigma_j^*\log\hat{\lambda}}{\sigma_j^*}\right)\right)^2}.$$

One can also verify that the densities of τ_j^* and $1/\sigma_j^*$ are both log-concave. So we also use the adaptive rejection sampling for drawing these unique pairs of location-scale parameters.

Next we update the configuration of clusters $\{c_i\}_{i=1}^n$, where c_i denotes the cluster indicator of the i-th observation. This is the key step of MCMC sampling for Dirichlet process mixture models. Note that in our case the base measure is not a conjugate prior. There have been many techniques developed for handling non-conjugate priors in MCMC sampling, for example [49, 77, 82, 109, 113]. We apply Algorithm 8 in [82] to sample efficiently using the Gibbs sampling with auxiliary parameters. Let l denote the number of auxiliary parameters. Let $n_{-i,c} := |\{j : 1 \le j \le n,\ j \ne i,\ c_j = c\}|$. For $i = 1, \cdots, n$, we iterate the following two steps.

(1) Let k^- be the number of distinct c_j's for $j \ne i$. Without loss of generality, we assume the set of distinct c_j's is $\{1, \cdots, k^-\}$. Let $\{(\tilde{\tau}_c, \tilde{\sigma}_c)\}_{c=1}^{k^-} = \{(\tau_c^*, \sigma_c^*)\}_{c=1}^{k^-}$ be the corresponding unique pairs. Set $h = k^- + l$.

- If $c_i = c_j$ for some $j \ne i$, draw independent samples from $\text{Logistic}(\tau| - \sigma\log\lambda, \sigma) \cdot \text{Inv-Gamma}(\sigma|c, d)$ for $(\tilde{\tau}_c, \tilde{\sigma}_c)$ where $k^- < c \le h$.
- If $c_i \ne c_j$ for all $j \ne i$, let $(\tilde{\tau}_{k^-+1}, \tilde{\sigma}_{k^-+1}) = (\tau_{c_i}^*, \sigma_{c_i}^*)$, and draw independent samples from $\text{Logistic}(\tau| - \sigma\log\lambda, \sigma) \cdot \text{Inv-Gamma}(\sigma|c, d)$ for $(\tilde{\tau}_c, \tilde{\sigma}_c)$ where $k^- + 1 < c \le h$.

Let $\tilde{\boldsymbol{\tau}} = (\tilde{\tau}_1, \cdots, \tilde{\tau}_h)^T$ and $\tilde{\boldsymbol{\sigma}} = (\tilde{\sigma}_1, \cdots, \tilde{\sigma}_h)^T$. Draw c_i from

$$P(c_i = c | n_{-i,c}, e_i, \tilde{\boldsymbol{\tau}}, \tilde{\boldsymbol{\sigma}}) = \begin{cases} \frac{K n_{-i,c}}{\tilde{\sigma}_c} \psi\left(\frac{e_i + \tilde{\tau}_c}{\tilde{\sigma}_c}\right), & \text{for } 1 \le c \le k^-; \\ \frac{K\alpha}{l} \frac{1}{\tilde{\sigma}_c} \psi\left(\frac{e_i + \tilde{\tau}_c}{\tilde{\sigma}_c}\right), & \text{for } k^- < c \le h; \end{cases}$$

where K is a normalizing constant.

(2) Update (τ_i, σ_i) by $(\tilde{\tau}_{c_i}, \tilde{\sigma}_{c_i})$.

We choose the number of auxiliary parameters to be 10, which balances the autocorrelation of samples and the computation time.

As discussed in Section 2.2, adjustment is required for the inference of the intercept. After each MCMC iteration, we need to compute q_F to perform the adjustment described in Proposition 2.2.1. Assume we have n^* clusters with cluster sizes s_j for $j = 1, \cdots, n^*$ and unique pairs $(\tau_j^*, \sigma_j^*)_{j=1}^{n^*}$. q_F is the p-th quantile of the mixture distribution with pdf $f(x) = \sum_{j=1}^{n^*} \frac{s_j}{n^*} \frac{1}{\sigma_j^*} \psi\left(\frac{x + \tau_j^*}{\sigma_j^*}\right)$ and is given by the equation

$$\int_{-\infty}^{q_F} \sum_{j=1}^{n^*} \frac{s_j}{n^*} \frac{1}{\sigma_j^*} \psi\left(\frac{x + \tau_j^*}{\sigma_j^*}\right) dx = p. \tag{2.8}$$

Because the logistic distribution has a closed-formed quantile function, we can rewrite (2.8) as

$$\sum_{j=1}^{n^*} \frac{s_j}{n^*} \frac{1}{1 + \frac{1-p}{p} \exp\left(-\frac{q_F + \tau_j^*}{\sigma_j^*}\right)} = p. \tag{2.9}$$

So we can solve for q_F from (2.9) by Newton's method. In addition, since $\min_j\{-\tau_j^*\} \le q_F \le \max_j\{-\tau_j^*\}$, we can also compute q_F quickly by a binary search.

In Step (iv), $\hat{\lambda}$ is the numerical solution to (2.3) in Proposition 2.1.1. And notice that one needs to compute $\hat{\lambda}$ only once for a given p. Although in general, rejection sampling may bring the concern of low computation efficiency due to rejections, the average number of rejection for each parameter is usually smaller than 2 in our experiment.

With the MCMC algorithm available for our model, we are able to perform extensive simulation studies to verify the correctness and computational advantage of our model.

2.4 Simulation study

In this section we conduct a systematic set of simulation experiments to compare our proposed DPML method with existing methods.

2.4.1 Ordinary designs

Six model designs with uncorrelated errors are used:

- **Design 1** : $Y_i = 1 + x_{1i}\beta_1 + x_{2i}\beta_2 + \epsilon_{1i}$;
- **Design 2** : $Y_i = 1 + x_{1i}\beta_1 + x_{2i}\beta_2 + \pi_i\epsilon_{1i} + (1-\pi_i)\epsilon_{2i}$;
- **Design 3** : $Y_i = 1 + x_{1i}\beta_1 + x_{2i}\beta_2 + \epsilon_{3i}$;
- **Design 4** : $Y_i = 1 + x_{3i}\beta_1 + (1.1 - x_{3i})\epsilon_{1i}$;
- **Design 5** : $Y_i = 1 + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{4i}\beta_3 + \epsilon_{1i}$;
- **Design 6** : $Y_i = 1 + x_{1i}\beta_1 + x_{2i}\beta_2 + \epsilon_{4i}$;

where $x_{1i}, x_{2i} \overset{iid}{\sim} N(0,1)$, $x_{3i} \overset{iid}{\sim}$ Uniform$(-1,1)$, $x_{4i} \overset{iid}{\sim} |t_2|$, $\epsilon_{1i} \overset{iid}{\sim} N(0,1)$, $\epsilon_{2i} \overset{iid}{\sim} N(3,3)$, $\epsilon_{3i} \overset{iid}{\sim}$ DoubleExp$(0,1)$, $\epsilon_{4i} \overset{iid}{\sim} t_3$ and $\pi_i \overset{iid}{\sim}$ Bernoulli(0.8). All covariates and error terms are mutually independent. We set $\beta_1 = \beta_2 = \beta_3 = 1$. Designs 1, 2, 3 and 6 are location shift models with different error distributions. Design 4 has heteroscedastic errors. Design 5 has a predictor x_4 from a heavy-tailed distribution, which violates the typical assumption in the frequentist asymptotic theory [61]. Designs 1-5 were previously used in [59, 89]. Design 6 adds a case for heavy-tailed error distribution.

For each design, 200 data sets each with sample size $n = 100$ are generated. And the quantiles of interest are $p = 0.5$ and 0.9. Each simulated data set is analyzed using the standard frequentist quantile regression method(FQR) implemented in the R [87] library **quantreg**, the Bayesian ALD method (BALD) implemented in the R library **bayesQR**, the DPM of uniform distributions (DPMU), the DPM of the mixtures of two normal distributions (DPMMN), our proposed DPM of logistic distributions (DPML). We implement all the above DPM-based methods using the R package **Rcpp**.

For all the Bayesian methods, after looking into convergence diagnostics such as trace plots and autocorrelation plots, we simulated 25,000 MCMC samples and used the first 5,000 as burn-in for each data set in each design. We set the thinning parameter to be 5. All methods are evaluated by the mean squared error (MSE) and predictive check loss (PCL),

$$\text{MSE} = \frac{1}{m}\sum_{i=1}^{m}\left(\beta_i - \hat{\beta}_i\right)^2 \quad \text{and} \quad \text{PCL} = \sum_{i=1}^{n}\rho_p(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}),$$

where m is the number of covariates with the intercept excluded, and β_i's are the true value, and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_m)^T$ are the estimate derived by taking the mean of the posterior distribution, and $\rho_p(u) = u[p - \mathbb{I}(u < 0)]$ is the check loss function. For each design, we generate a data set of size 10,000 as the validation data set, based on which we calculate the predictive check loss.

For our DPML model in (2.1), we set the hyper-parameters $a_1 = b_1 = 1$. In general, posterior predictive inference is robust to the prior choice for the concentration parameter α, as claimed in [66] and shown through simulation in [89]. Also we remark that it is possible to incorporate the idea of empirical Bayes to estimate the concentration parameter as in [79]. We set $c = 2$, $\nu = 10^8$ and for a_2 and b_2, we use the empirical Bayes method. Note that for the prior of σ_i's, they all have mean equal to d. We want to make

the prior mean of d large enough such that the dispersion in the data can be captured. To achieve this we simply set $a_2 = 1$ and $b_2 = \max_{i=1,\cdots,n} y_i - \min_{i=1,\cdots,n} y_i$.

The hyperparameters in DPMMN and DPMU are set similarly as in DPML. Hyperparameters related to the scale parameters of the kernel densities are specified by the empirical Bayes method, and the prior variance of the regression coefficients are set to be large. Note here due to the complexity of DPMMN, we follow [89] to approximate the DPMMN by a finite mixture models with the number of components equal to 10. And in our simulation study, we find that further increasing the number of components only increase the computational time and has no significant improvement in the performance in terms of inference.

For each method and each design, we report the mean and standard error (for the mean) of 200 MSE's, average coverage probability of the 90% credible or confidence interval for the slope parameters and the mean and standard error of the PCL's. The results are summarized in Tables 2.1, 2.2 and 2.3. In Table 2.1, all the reported values are 100×average (100×standard error) over the 200 simulated data sets. In Table 2.2, the reported values are $\sum_{i=1}^{m} c_i/m$ and $\sum_{i=1}^{m} l_i/m$, where l_i denotes the average length of 90% credible intervals of $\hat{\beta}_i$ and c_i denotes the empirical coverage probability of the 90% credible interval of $\hat{\beta}_i$. We also get the mean squared error and coverage probability for each regression coefficients including the intercept. All the detailed results can be find in Appendix A.3.

In terms of the MSE of the regression coefficient estimates, DPMMN and our proposed method DPML outperform other methods for both $p = 0.5$ and $p = 0.9$ in Designs 1-3, 5 and 6, except that in Design 3, for $p = 0.5$, BALD has the best performance. And this is the case where the data are generated from the model assumed by the BALD method.

However, at the more extreme quantile $p = 0.9$, DPML and DPMMN still significantly outperform BALD. Design 4 involves heteroscedasticity and all DPM-based methods have very poor performance, especially for $p = 0.9$.

In terms of coverage probability, DPML, DPMMN and FQR all have coverage probabilities close to the nominal level 0.9 when there is no heteroscedasticity, i.e. Designs 1-3, 5, 6. However, FQR usually has longer intervals than the Bayesian competitors. But in the presence of heteroscedasticity as in Design 4, DPM-based methods all have poor coverage probabilities which again shows the need for explicitly modelling the heteroscedasticity to be discussed in Chapter 4.

Based on the MSE and coverage probability, there seems to be no significant difference in the performance between DPML and DPMMN. However, DPMMN is essentially based on a DPM over four parameters while our method only mixes over two parameters. Our method is more appealing in terms of computational efficiency as discussed in Section 2.4.3.

As for the predictive check loss, DPML and DPMMN have the best performance when there is no heteroscedasticity, and DPM-based methods do not perform as well as FQR or BALD in the presence of heteroscedasticity. For $p = 0.9$, in Design 2, DPMMN performs significantly better in DPML, while in Design 6 DPML performs significantly better. This is as expected, since Design 2 is the same as the model specification of DPMMN and Design 6 has a heavy-tailed error distribution, which is better handled by our logistic kernel.

2.4.2 Robustness to outliers

One important motivation to use the logistic kernel density is to achieve robustness against outliers, due to its heavier tail than the normal distribution. We focus on outliers that are data generated from distributions that violate the quantile constraint.

We consider two scenarios with contamination proportions 5% and 10%, respectively. Again the quantiles of interest are 0.5 and 0.9.

- $p = 0.5$, $y_i = 1 + x_{1i}\beta_1 + x_{2i}\beta_2 + \pi_i\epsilon_{1i} + (1-\pi_i)\epsilon_{2i}$;
- $p = 0.9$, $y_i = 1 + x_{1i}\beta_1 + x_{2i}\beta_2 + \pi_i(\epsilon_{1i} - qnorm(0.9)) + (1-\pi_i)\epsilon_{2i}$,

where $\epsilon_{1i} \overset{iid}{\sim} N(0,1)$, $\epsilon_{2i} \overset{iid}{\sim} N(5, 0.01)$ and $\pi_i \overset{iid}{\sim} Bernoulli(0.95)$ for 5% contamination and $\pi_i \overset{iid}{\sim} Bernoulli(0.9)$ for 10% contamination. We set $\beta_1 = \beta_2 = 1$.

We generated 200 data sets each of size 500 under each combination of the contamination proportion and the quantile of interest. We still compare the performance in terms of MSE of the regression coefficients, the coverage probability of 90% credible or confidence intervals and the PCL. The PCL is evaluated on validation data sets that has no contamination. Results are summarized in Tables 2.4, 2.5 and 2.6. The details about the intercept estimation and the simulation result for a smaller sample size $n = 100$ can be found in Appendix A.3.

In terms of the MSE of the regression coefficients, DPML outperforms DPMMN for both scenarios and both quantiles. The difference is statistically significant. As for the 90% credible intervals, DPML usually have coverage probability closer to 0.9 and have shorter intervals compared to DPMMN. And for the predictive check loss, DPML performs better than DPMMN for most cases except for the extreme quantile $p = 0.9$ with 10% of outliers. This is due to the fact that 10% of outliers make the estimation of

the intercept totally unreliable while the intercept has a big impact on the calculation of the check loss. And both methods perform very poorly.

Overall, by choosing a heavy-tailed kernel with heavier tail we can achieve robustness against outliers.

2.4.3 Computational efficiency

In this section, we compare the computational efficiency between DPML and DPMMN. To make a fair comparison, both methods use the stick-breaking representation (1.4) with 10 components to simulate the Dirichlet process. All the parameters in DPML are updated by the Metropolis-Hastings algorithm. The regression coefficients of DPMMN are updated by Gibbs sampling and all the other parameters in DPMMN are updated by the Metropolis-Hastings algorithm. We generated data sets from Designs 1 and 6 with $n = 100k$, $k = 1, \cdots, 20$, and set $p = 0.5$. For each design and each sample size, we perform quantile regressions using DPML and DPMMN.

The results are summarized in Figures 2.4, 2.5 and 2.6. From Figure 2.4, it is clear that using our logistic kernel significantly reduces the computation time. And since logistic distribution has heavier tail than the normal distribution, our DPML model has a much smaller number of components (clusters) empirically, especially when the true distribution is heavy-tailed as in Design 6. The comparison on the numbers of clusters and the numbers of clusters with proportion larger the 5% is shown in Figure 2.5 and Figure 2.6, respectively. Practically, in DPM models, the computational efficiency depends on the number of latent parameters, which is in turn determined by the number of clusters and the number of parameters in the kernel. The DPML model has a simpler

kernel with two parameters and fewer clusters compared to the DPMMN model, thus the DPML model is more appealing.

Figure 2.4. Comparison of the computational efficiency between DPML and DPMMN with $p = 0.5$.

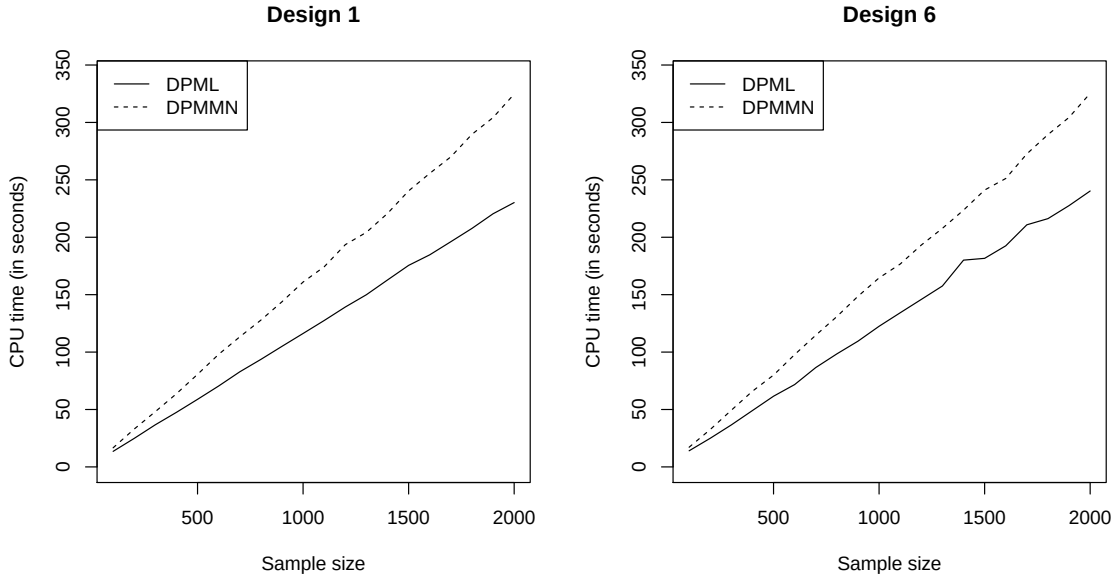


Figure 2.5. Comparison of the numbers of clusters.

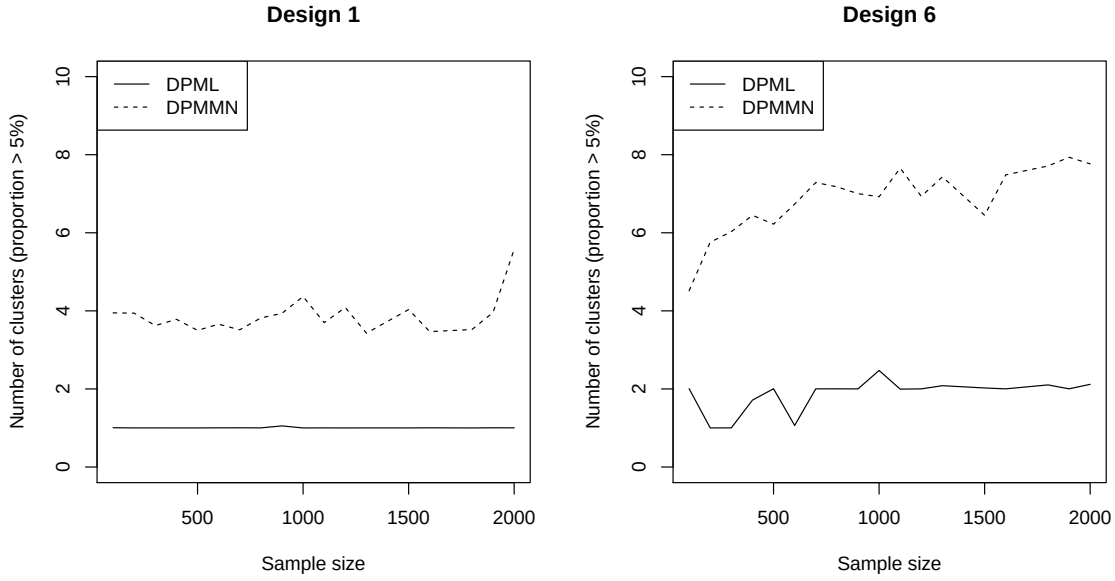
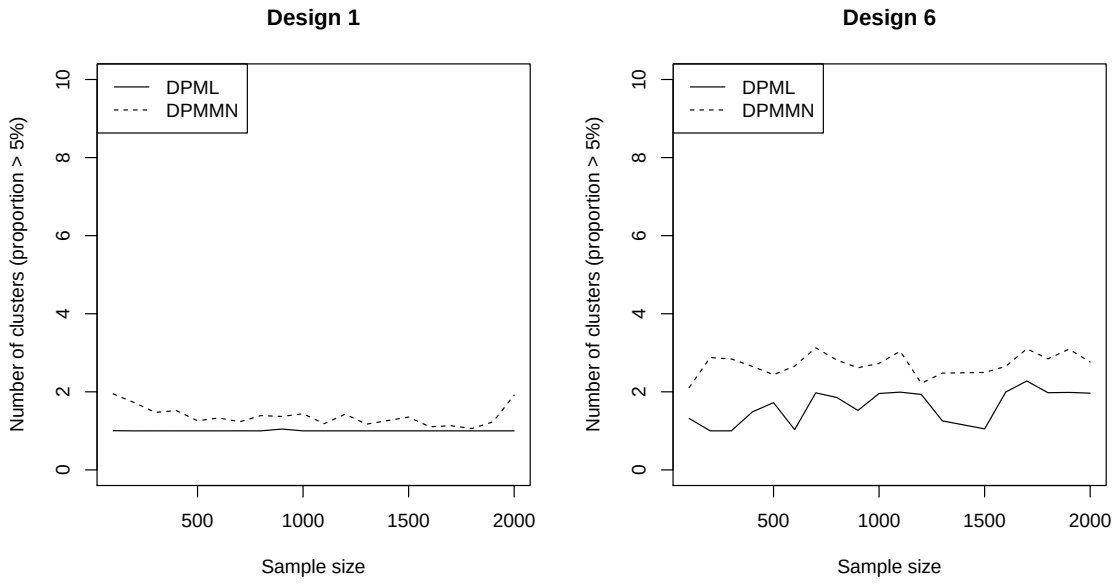


Figure 2.6. Comparison of the numbers of clusters with proportion larger the 5%.



Remark 2.1

In case the sample size is very large or the data have high dimensionality, the MCMC algorithms for DPM models can be very slow, because the number of latent parameters, e.g. unique pairs (τ_j^*, σ_j^*) in DPML, increases with the sample size. Some efforts have been devoted to develop methods to speed up the inference of DPM models.

Within the MCMC framework, we can increase the computational efficiency by truncating the DPM model to a finite mixture model as in [89]. However, the number of components should be large enough to avoid sacrificing too much accuracy. As the sample size increases, the required number of components also gets larger. Therefore, MCMC algorithms for the truncated DPM models are still infeasible for very large data sets. There are also approaches to parallelize the MCMC algorithms for DPM models for efficient computation in distributed systems [19, 75, 76, 115]. These methods rely on the conditional independence of observations between clusters. One bottleneck of these

parallel computing methods is that, for each iteration of the MCMC sampling we have to shuffle the data among different machines in the distributed system, which is usually computationally expensive. But parallel MCMC algorithms do make it possible to analyze extreme large data sets using DPM models.

An alternative to MCMC methods is the variational inference methods [85,106], which turn the computation of posterior distribution into an optimization problem. Given a posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$, the variational method approximates $p(\boldsymbol{\theta}|\boldsymbol{x})$ by the so-called variational distribution $q(\boldsymbol{\theta}|\boldsymbol{\nu})$ with a known form, where $\boldsymbol{\nu}$ is the so-called variational parameter. The goal is to find

$$\hat{\boldsymbol{\nu}} = \arg\min_{\boldsymbol{\nu}} D_{KL}\left[q(\boldsymbol{\theta}|\boldsymbol{\nu})||p(\boldsymbol{\theta}|\boldsymbol{x})\right],$$

where D_{KL} is the Kullback-Leibler divergence, that is, $D_{KL}(f||g) = \int f(x)\log\frac{f(x)}{g(x)}dx$. The key of variational inference is to find $q(\boldsymbol{\theta}|\boldsymbol{\nu})$ such the optimization problem is easy to solve. [10] develops a variational inference method for DPM models by specifying $q(\boldsymbol{\theta}|\boldsymbol{\nu})$ as a truncated DPM. Although the proposed method can substantially speed up the posterior inference for the DPM models, it is only applicable to the case where the kernel is conjugate to the base measure. Without the conjugacy, one has to solve an optimization problem within each step of the iterative procedure to solve the initial optimization problem. This increases the computation burden. Another drawback of the variational methods is that the optimization procedure suffers from local maxima in the variational parameter space. Finally, there is no theory to evaluate the approximation error. Despite the disadvantages of the variational methods, they are still very popular in practice due to their sampling-free nature.

Sequential updating algorithms are also proposed to accelerate the posterior inference of DPM models. [83] develops a recursive estimation algorithm to approximate the

Bayesian density estimate under a DPM prior. The algorithm is motivated by the full conditional distribution of the latent parameters, and observations are sequentially utilized to update the estimate. The posterior consistency is established under some restrictions in [43, 104]. [111] proposes a sequential greedy search algorithm for selecting the cluster for each observation in DPM models. The idea is to view the grouping of observations into clusters as a model selection problem. Both sequential updating algorithms [83, 111] are very fast in computation. However, they both require the kernel density to be conjugate to the base measure, and the estimates of both methods depends on the ordering of the observations. Besides, both methods are mainly developed for the purpose of density estimation and difficult to extend to the regression problems.

Due to the computational challenge of applying DPM models to large data sets, it is desirable to choose simple kernels which are also conjugate to the base measure. The conjugacy makes efficient computation methods such as the variational inference methods and sequential updating algorithms applicable. With this motivation, a DPM model with the normal distribution as the kernel is proposed in Chapter 5.

Table 2.1
Average MSE for the regression coefficients (except the intercept) for $p = 0.5$ and $p = 0.9$. MSE is reported as 100 × average (100 × standard error) over the 200 simulated data sets for each design.

p	Design	FQR	BALD	DPMU	DPMMN	DPML
0.5	1	1.58	1.30	1.45	**1.06**	**1.06**
		(0.11)	(0.10)	(0.11)	(0.08)	(0.08)
0.5	2	2.45	2.15	2.17	1.87	**1.72**
		(0.18)	(0.16)	(0.15)	(0.13)	(0.12)
0.5	3	1.61	**1.49**	2.15	1.56	1.67
		(0.15)	(0.13)	(0.19)	(0.14)	(0.15)
0.5	4	4.48	**4.39**	18.85	5.98	9.07
		(0.46)	(0.46)	(1.54)	(0.63)	(0.93)
0.5	5	1.31	1.08	1.16	**0.88**	0.90
		(0.10)	(0.07)	(0.08)	(0.05)	(0.05)
0.5	6	2.27	1.96	2.51	1.90	**1.87**
		(0.16)	(0.14)	(0.19)	(0.13)	(0.13)
0.9	1	3.35	2.76	1.51	**1.06**	**1.06**
		(0.27)	(0.22)	(0.12)	(0.08)	(0.07)
0.9	2	27.24	20.76	3.13	**1.66**	1.76
		(1.66)	(1.25)	(0.25)	(0.11)	(0.12)
0.9	3	8.84	7.22	2.66	**1.69**	1.70
		(0.67)	(0.59)	(0.21)	(0.15)	(0.15)
0.9	4	9.22	**8.27**	124.40	177.52	160.60
		(0.99)	(0.89)	(6.82)	(7.01)	(5.72)
0.9	5	2.57	1.91	1.21	0.91	**0.90**
		(0.16)	(0.13)	(0.08)	(0.05)	(0.05)
0.9	6	9.86	8.07	2.75	1.96	**1.85**
		(0.73)	(0.60)	(0.20)	(0.14)	(0.13)

Table 2.2

Average coverage probabilities (CP) of 90% credible or confidence intervals of the regression coefficients (except the intercept) for $p = 0.5$ and $p = 0.9$. The average length of intervals is also reported.

p	Design		FQR	BALD	DPMU	DPMMN	DPML
0.5	1	CP	0.89	0.82	0.81	**0.90**	0.88
		Length	0.42	0.34	0.32	0.34	0.34
0.5	2	CP	0.89	0.89	0.83	**0.90**	0.91
		Length	0.54	0.48	0.42	0.47	0.45
0.5	3	CP	0.88	0.88	0.83	**0.90**	**0.90**
		Length	0.40	0.37	0.39	0.39	0.40
0.5	4	CP	0.80	0.79	0.52	0.80	0.76
		Length	0.58	0.56	0.80	0.65	0.77
0.5	5	CP	**0.89**	0.82	0.84	**0.89**	0.88
		Length	0.35*	0.28	0.28	0.29	0.29
0.5	6	CP	0.87	0.84	0.81	0.88	**0.89**
		Length	0.45	0.41	0.42	0.43	0.43
0.9	1	CP	0.88	0.67	0.81	**0.89**	0.88
		Length	0.57	0.33	0.32	0.34	0.34
0.9	2	CP	**0.89**	0.56	0.77	0.92	0.92
		Length	1.53	0.78	0.42	0.45	0.46
0.9	3	CP	0.87	0.64	0.75	0.89	**0.90**
		Length	0.92	0.52	0.38	0.40	0.41
0.9	4	CP	0.81	0.64	0.03	0.01	0.01
		Length	0.92	0.56	0.76	0.75	0.74
0.9	5	CP	**0.89**	0.70	0.83	**0.89**	0.88
		Length	0.71	0.29	0.28	0.30	0.29
0.9	6	CP	0.88	0.61	0.79	0.87	**0.89**
		Length	0.92	0.53	0.42	0.43	0.43

*Since x_{4i} is generated from t_2, FQR produces confidence interval with infinite length for some simulated data sets. As a result, the average length of the confidence interval is infinite. For comparison, we use the median of the lengths of the frequentist confidence interval. For other methods, there is little difference between the mean and the median of the lengths of credible intervals.

Table 2.3
Predictive check loss for $p = 0.5$ and $p = 0.9$. Standard error is reported in the parenthesis.

p	Design	FQR	BALD	DPMU	DPMMN	DPML
0.5	1	4147.1	4133.48	4140.55	**4118.53**	4120.07
		(5.28)	(4.47)	(5.4)	(3.58)	(3.67)
0.5	2	6249.45	6234.25	6232.86	6226.86	**6216.28**
		(7.31)	(6.61)	(6.72)	(6.29)	(5.77)
0.5	3	5098.31	5091.85	5118.82	**5097.82**	5102.46
		(6.7)	(6.15)	(7.58)	(6.35)	(6.67)
0.5	4	4372.87	**4368.4**	4489.24	4379.28	4404.81
		(5.72)	(5.38)	(12.97)	(6.32)	(8.44)
0.5	5	4149.59	4123.11	4132.29	**4101.74**	4103.67
		(8.28)	(6.46)	(7.48)	(5.32)	(5.41)
0.5	6	5597.85	5580.9	5600.52	5574.02	**5573.75**
		(6.68)	(5.73)	(7.14)	(5.36)	(5.33)
0.9	1	1852.52	1837.36	1843.54	**1802.19**	1802.49
		(5.4)	(4.53)	(5.42)	(2.45)	(2.7)
0.9	2	4052.86	3996.39	4254.92	**3877.11**	3902.59
		(12.4)	(9.61)	(21.15)	(6.12)	(8.28)
0.9	3	2718.23	2694.21	2764.1	2632.21	**2625.03**
		(8.41)	(7.31)	(11.31)	(3.46)	(3.03)
0.9	4	1972.84	**1968.46**	2329.13	2357.3	2327.29
		(4.34)	(3.76)	(18.95)	(13.47)	(12.27)
0.9	5	1924.57	1879.18	1878.67	**1831.51**	1833.25
		(9.29)	(7.36)	(6.98)	(4.51)	(4.53)
0.9	6	3048.02	3022.22	3032.46	2965.25	**2949.79**
		(8.58)	(7.18)	(8.39)	(4)	(2.86)

Table 2.4

MSE for the regression coefficients when there are outliers. MSE is reported as 100 × average (100 × standard error) over the 200 simulated data sets.

Contamination proportion	p	Coefficient	DPMMN	DPML
5%	0.5	β_1	0.073	**0.039**
			(0.009)	(0.004)
5%	0.5	β_2	0.067	**0.039**
			(0.007)	(0.004)
5%	0.9	β_1	0.06	**0.037**
			(0.006)	(0.004)
5%	0.9	β_2	0.075	**0.048**
			(0.008)	(0.005)
10%	0.5	β_1	0.028	**0.019**
			(0.002)	(0.002)
10%	0.5	β_2	0.037	**0.021**
			(0.009)	(0.002)
10%	0.9	β_1	0.033	**0.022**
			(0.004)	(0.002)
10%	0.9	β_2	0.031	**0.02**
			(0.003)	(0.002)

Table 2.5
Lengths of 90% credible or confidence intervals and coverage probabilities (CP) for regression coefficients when there are outliers.

Contamination proportion	p	Coefficient		DPMMN	DPML
5%	0.5	β_1	CP	0.985	**0.915**
			Length	0.109	0.069
5%	0.5	β_2	CP	0.965	**0.92**
			Length	0.109	0.069
5%	0.9	β_1	CP	0.98	**0.92**
			Length	0.107	0.069
5%	0.9	β_2	CP	0.95	**0.905**
			Length	0.108	0.069
10%	0.5	β_1	CP	1	**0.91**
			Length	0.084	0.047
10%	0.5	β_2	CP	0.97	**0.905**
			Length	0.085	0.049
10%	0.9	β_1	CP	0.98	**0.88**
			Length	0.085	0.047
10%	0.9	β_2	CP	0.975	**0.915**
			Length	0.084	0.047

Table 2.6
Predictive check loss when there are outliers. The standard error is reported in the parenthesis.

Contamination proportion	p	DPMMN	DPML
5%	0.5	3974.75	**3973.01**
		(1.098)	(1.047)
5%	0.9	2353.03	**1887.64**
		(77.859)	(2.553)
10%	0.5	4013.48	**4008.68**
		(2.279)	(2.07)
10%	0.9	**5225.74**	5748.72
		(44.677)	(4.04)

3. Posterior consistency

In this chapter, we prove the posterior consistency of our proposed DPML model in Chapter 2. As shown in Section 2.2, with the adjustment described in Proposition 2.2.1, our model (2.1) is practically equivalent to a DPM model (2.5) which employs a location shift of the mixture to make the random probability density satisfy the quantile constraint. Therefore, it is sufficient to show the posterior consistency of the model (2.5).

Let $\mathcal{F}_p$ denote the space of probability densities that have their p-th quantile equal to 0. Let $(\psi * P)(x) := \int \frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right) dP(\tau, \sigma)$ for a given mixing distribution P and define q_P as the p-th quantile of $(\psi * P)(x)$. We obtain a prior Π^* over $\mathcal{F}_p$ by defining the random probability density $f(x) = (\psi * P)(x - q_P)$, where $P \sim DP(\alpha, G)$, and still let $\tilde{\Pi}$ denote $DP(\alpha, G)$.

Definition 3.0.1 *Given a prior* Π *over* $\mathcal{F}_p$*, let* $f_0 \in \mathcal{F}_p$ *be the true density. We say that* Π *has the Kullback-Leibler property if* $\Pi(K_\epsilon(f_0)) > 0$ *for all* $\epsilon > 0$*, where* $K_\epsilon(f) := \{g : K(f, g) < \epsilon\}$ *and* $K(f, g) := \int f(x) \log \frac{f(x)}{g(x)} dx$.

The Kullback-Leibler property plays a fundamental role in the posterior consistency theory. We will derive the Kullback-Leibler property for our prior Π^* in Section 3.1 and show the posterior consistency in Section 3.2.

3.1 Kullback-Leibler property

We first provide an important lemma characterizing the tail behaviour of q_P, which is the p-th quantile of $(\psi * P)(x)$ with $P \sim DP(\alpha, G)$, where G is a probability measure over $\mathbb{R} \times \mathbb{R}^+$.

Lemma 3.1.1 *Assume that there exist $\gamma > 0$ and $c > 0$ such that for large $x > 0$,*

$$G((-\infty, x) \times (0, x)) > 1 - cx^{-\gamma} \text{ and } G((-x, \infty) \times (0, x)) > 1 - cx^{-\gamma}.$$

Then there exists a constant $d > 0$ such that for any $t > 0$,

$$Pr(|q_P| \leq t) > 1 - \frac{d}{t^\gamma}.$$

Proof Pick any $0 < \epsilon < t$ and $\nu > 0$. And define

$$a_{\epsilon,\nu} := \max\left\{ p\left(1 + \frac{1-p}{p}\exp\left(-\frac{\epsilon}{\nu}\right)\right), 1 - p + \frac{1}{1 + \frac{1-p}{p}\exp\left(\frac{\epsilon}{\nu}\right)} \right\}.$$

It is easy to verify that $a_{\epsilon,\nu} < 1$. Also let $S_{t,\epsilon,\nu} := [\epsilon - t, t - \epsilon] \times (0, \nu)$. Then given any probability measure P over $\mathbb{R} \times \mathbb{R}^+$ if $P\left(S_{t,\epsilon,\nu}\right) \geq a_{\epsilon,\nu}$, we have

$$\begin{aligned}
&\int_{-\infty}^{-t} (\psi * P)(x)dx = \int \frac{1}{1 + \frac{1-p}{p}\exp\left(\frac{\tau+t}{\sigma}\right)} dP(\tau, \sigma) \\
=&\int_{S_{t,\epsilon,\nu}} \frac{1}{1 + \frac{1-p}{p}\exp\left(\frac{\tau+t}{\sigma}\right)} dP(\tau, \sigma) + \int_{S^c_{t,\epsilon,\nu}} \frac{1}{1 + \frac{1-p}{p}\exp\left(\frac{\tau+t}{\sigma}\right)} dP(\tau, \sigma) \\
\leq&\frac{1}{1 + \frac{1-p}{p}\exp\left(\frac{\epsilon}{\nu}\right)} P\left(S_{t,\epsilon,\nu}\right) + P\left(S^c_{t,\epsilon,\nu}\right) \\
\leq&\frac{1}{1 + \frac{1-p}{p}\exp\left(\frac{\epsilon}{\nu}\right)} + 1 - a_{\epsilon,\nu} \leq p,
\end{aligned} \tag{3.1}$$

and

$$\begin{aligned}
&\int_{-\infty}^{t} (\psi * P)(x)dx = \int \frac{1}{1 + \frac{1-p}{p}\exp\left(\frac{\tau-t}{\sigma}\right)} dP(\tau, \sigma) \\
\geq&\int_{S_{t,\epsilon,\nu}} \frac{1}{1 + \frac{1-p}{p}\exp\left(\frac{\tau-t}{\sigma}\right)} dP(\tau, \sigma) \geq \frac{1}{1 + \frac{1-p}{p}\exp\left(-\frac{\epsilon}{\nu}\right)} P\left(S_{t,\epsilon,\nu}\right) \geq p.
\end{aligned} \tag{3.2}$$

By equations (3.1) and (3.2), we have $|q_P| \leq t$. Therefore,

$$Pr(|q_P| \leq t) \geq Pr[P\left(S_{t,\epsilon,\nu}\right) \geq a_{\epsilon,\nu}] = Pr[1 - P\left(S_{t,\epsilon,\nu}\right) \leq 1 - a_{\epsilon,\nu}]. \tag{3.3}$$

Since $P \sim DP(\alpha, G)$, $1 - P\left(S_{t,\epsilon,\nu}\right) \sim$ Beta$(\alpha G(S^c_{t,\epsilon,\nu}), \alpha G(S_{t,\epsilon,\nu}))$. Therefore, applying the Markov inequality, we have

$$\begin{aligned} &Pr[1 - P\left(S_{t,\epsilon,\nu}\right) \leq 1 - a_{\epsilon,\nu}] = 1 - Pr[1 - P\left(S_{t,\epsilon,\nu}\right) \geq 1 - a_{\epsilon,\nu}] \\ \geq &1 - \frac{E[1 - P\left(S_{t,\epsilon,\nu}\right)]}{1 - a_{\epsilon,\nu}} = 1 - \frac{1 - G(S_{t,\epsilon,\nu})}{1 - a_{\epsilon,\nu}}. \end{aligned} \tag{3.4}$$

Since we are free to choose any $0 < \epsilon < t$ and $\nu > 0$, we now set $\epsilon = \nu = \frac{t}{2}$. Then $a := a_{\epsilon,\nu}$ is a constant which does not depend on t, and $S_{t,\epsilon,\nu} = \left[-\frac{t}{2}, \frac{t}{2}\right] \times \left(0, \frac{t}{2}\right)$. Now by the assumption on the tail of G, $G(S_{t,\epsilon,\nu}) > 1 - 2ct^{-\gamma}$. Plug $G(S_{t,\epsilon,\nu})$ and a into equation (3.4), we get

$$Pr(1 - P\left(S_{t,\epsilon,\nu}\right) \leq 1 - a_{\epsilon,\nu}) >= 1 - \frac{2c}{(1-a)t^{\gamma}}. \tag{3.5}$$

The result follows by combining equation (3.5) with equation (3.3). ■

Now we state a lemma similar to Lemma 3.2 in [102], which is the main tool to show the Kullback-Leibler property of f_0. For any $h > 0$, define a probability measure P_0^h over $\mathbb{R} \times \mathbb{R}^+$ such that $dP_0^h = f_0 \times \delta_h$.

Lemma 3.1.2 *Suppose $f_0 \in \mathcal{F}_p$, the space of probability densities whose p-th quantiles equal to 0. If for any $0 < \delta < 1$ and $\gamma > 0$, there exist a set $\mathcal{A}$ and a constant $x_0 > 0$ such that $\tilde{\Pi}(\mathcal{A}) > 1 - \delta$ and $\int_{|x|>x_0} f_0(x) \log \frac{f_0(x)}{f(x)} dx < \gamma$, for any $f(x) = (\psi * P)(x - q_P)$ with $P \in \mathcal{A}$.*

And assume that there exist $\eta > 0$ and $c > 0$ such that for large $x > 0$,

$G((-\infty, x) \times (0, x)) > 1 - cx^{-\eta}$ *and* $G((-x, \infty) \times (0, x)) > 1 - cx^{-\eta}$.

Also assume that

$(C1)$ $\lim_{h\to 0} \int_{-x_0}^{x_0} f_0(x) \log \frac{f_0(x)}{\int_{-x_0}^{x_0} \frac{1}{h}\psi\left(\frac{x-\tau}{h}\right) f_0(\tau)d\tau} dx = 0.$

Further assume that

$(C2)$ *there exists* $h_0 > 0$ *such that for any* $h < h_0$, P_0^h *is in the weak support of* $\tilde{\Pi}$.

Then Π^* *possesses the Kullback-Leibler property.*

Proof The proof of this lemma uses the same approach as in [102], which relies on results developed in [38].

Firstly, by the assumption, fix $0 < \delta < 1$ and $\epsilon > 0$, pick $x_0 > z_0$ and $\mathcal{A}$ such that $\tilde{\Pi}(\mathcal{A}) > 1 - \delta$. Then for any $f = \psi * P$ with $P \in \mathcal{A}$,

$$\int_{|x|>x_0} f_0(x) \log \frac{f_0(x)}{f(x)} dx < \frac{\epsilon}{2}. \tag{3.6}$$

Secondly, by assumption $(C1)$, there exists $\sigma_1 \in (0, h_0)$, such that

$$\int_{-x_0}^{x_0} f_0(x) \log \frac{f_0(x)}{\int_{-x_0}^{x_0} \frac{1}{\sigma_1}\psi\left(\frac{x-\tau}{\sigma_1}\right) f_0(\tau)d\tau} dx < \frac{\epsilon}{6}. \tag{3.7}$$

Thirdly, fix any $\kappa > 0$ and pick $0 < \lambda < 1$ such that $1 - \frac{\lambda}{\kappa^2(1-\lambda)} > \delta$. Choose a compact set $K \subset \mathbb{R} \times \mathbb{R}^+$ large enough such that $[-x_0, x_0] \times \{\sigma_1\} \subset K$, $G(K) > 1-\lambda$ and $P_0^{\sigma_1}(K) > 1-\lambda$. Let $\mathcal{E} := \left\{P : \left|\frac{P(K)}{P_0^{\sigma_1}(K)} - 1\right| < \kappa\right\}$. Since $P(K) \sim \text{Beta}(\alpha G(K), \alpha G(K^c)$, we have

$$\begin{aligned} E\left[(P(K) - P_0^{\sigma_1}(K))^2\right] &= E\left[(P(K) - G(K) + G(K) - P_0^{\sigma_1}(K))^2\right] \\ &= Var(P(K)) + (G(K) - P_0^{\sigma_1}(K))^2 \leq \lambda + \lambda^2. \end{aligned}$$

By applying Chebyshev's inequality, we get

$$\tilde{\Pi}(\mathcal{E}) \geq 1 - \frac{E\left[(P(K) - P_0^{\sigma_1}(K))^2\right]}{\kappa^2 P_0^{\sigma_1}(K)^2} \geq 1 - \frac{\lambda + \lambda^2}{\kappa^2(1-\lambda)^2} > \delta.$$

Therefore, $\tilde{\Pi}(\mathcal{A} \cap \mathcal{E}) > 0$.

Define the conditional probability under P given K by $P^*(S) := \frac{P(S \cap K)}{P(K)}$ for any $S \subseteq$

$\mathbb{R}\times\mathbb{R}^+$. Similarly, define $P_0^{\sigma_1*}(S):=\frac{P_0^{\sigma_1}(S\cap K)}{P_0^{\sigma_1}(K)}$. Let $c:=\inf\limits_{|x|<x_0}\inf\limits_{(\tau,\sigma)\in K}\frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right)$. Clearly $c>0$. Let

$$\mathcal{G}:=\left\{P:\ \sup_{-x_0\le x\le x_0}\log\left(\frac{\int\limits_K\frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right)dP_0^{\sigma_1*}(\tau,\sigma)}{\int\limits_K\frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right)dP^*(\tau,\sigma)}\right)<3\kappa+\frac{\kappa}{c}\right\}.$$

We claim $\tilde{\Pi}(\mathcal{A}\cap\mathcal{E}\cap\mathcal{G})>0$. The proof is given in Lemma 3.1.3. For $P\in\mathcal{E}\cap\mathcal{G}$, by choosing a proper κ, we have

$$\begin{aligned}
&\int_{-x_0}^{x_0}f_0(x)\log\frac{\int_K\frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right)dP_0^{\sigma_1}(\tau,\sigma)}{\int_K\frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right)dP(\tau,\sigma)}dx\\
=&\int_{-x_0}^{x_0}f_0(x)\left(\log\frac{\int_K\frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right)dP_0^{\sigma_1*}(\tau,\sigma)}{\int_K\frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right)dP^*(\tau,\sigma)}+\log\frac{P_0^{\sigma_1}(K)}{P(K)}\right)dx\\
\le&\sup_{-x_0\le x\le x_0}\log\left(\frac{\int\limits_K\frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right)dP_0^{\sigma_1*}(\tau,\sigma)}{\int\limits_K\frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right)dP^*(\tau,\sigma)}\right)+\frac{P_0^{\sigma_1}(K)}{P(K)}-1\\
\le&3\kappa+\frac{\kappa}{c}+\frac{\kappa}{1-\kappa}<\frac{\epsilon}{6}.
\end{aligned}\tag{3.8}$$

Fourthly, pick any $M>0$, let $d:=\inf\limits_{|x|<x_0+M}\inf\limits_{(\tau,\sigma)\in K}\frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right)$. Clearly $d>0$. Since K is compact, there exists $t_0\in(0,M)$, such that for $|t|<t_0$,

$$\sup_{|x|<x_0,(\tau,\sigma)\in K}\left|\frac{1}{\sigma}\psi\left(\frac{x-\tau\pm t}{\sigma}\right)-\frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right)\right|<\frac{d\epsilon}{6}.$$

Let $\mathcal{J}:=\{P:\ |q_P|<t_0\}$, then by Lemma 3.1.1, $\tilde{\Pi}(\mathcal{J})>0$. We will show in Lemma 3.1.4 that $\tilde{\Pi}(\mathcal{A}\cap\mathcal{E}\cap\mathcal{G}\cap\mathcal{J})>0$. If $P\in\mathcal{J}$,

$$\begin{aligned}
&\int_{-x_0}^{x_0}f_0(x)\log\frac{\int_K\frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right)dP(\tau,\sigma)}{\int_K\frac{1}{\sigma}\psi\left(\frac{x-\tau-q_P}{\sigma}\right)dP(\tau,\sigma)}dx\\
\le&\int_{-x_0}^{x_0}f_0(x)\left(\frac{\int_K\frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right)dP(\tau,\sigma)}{\int_K\frac{1}{\sigma}\psi\left(\frac{x-\tau-q_P}{\sigma}\right)dP(\tau,\sigma)}-1\right)dx\\
\le&\int_{-x_0}^{x_0}f_0(x)\frac{\left|\int_K\frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right)dP(\tau,\sigma)-\int_K\frac{1}{\sigma}\psi\left(\frac{x-\tau-q_P}{\sigma}\right)dP(\tau,\sigma)\right|}{d}dx\\
<&\frac{\epsilon}{6}.
\end{aligned}\tag{3.9}$$

If $P \in \mathcal{A} \cap \mathcal{E} \cap \mathcal{G} \cap \mathcal{J}$, by (3.6), (3.7), (3.8) and (3.9), we have

$$
\begin{aligned}
&\int f_0(x) \log \frac{f_0(x)}{f(x)} dx \\
\leq &\int_{-x_0}^{x_0} f_0(x) \log \frac{f_0(x)}{\int_K \frac{1}{\sigma}\psi\left(\frac{x-\tau-q_P}{\sigma}\right) dP(\tau,\sigma)} + \int_{|x|>x_0} f_0(x) \log \frac{f_0(x)}{f(x)} dx \\
\leq &\int_{-x_0}^{x_0} f_0(x) \log \frac{f_0(x)}{\int_{-x_0}^{x_0} \frac{1}{\sigma_1}\psi\left(\frac{x-\tau}{\sigma_1}\right) f_0(\tau) d\tau} dx \\
&+ \int_{-x_0}^{x_0} f_0(x) \log \frac{\int_K \frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right) dP_0^{\sigma_1}(\tau,\sigma)}{\int_K \frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right) dP(\tau,\sigma)} dx \\
&+ \int_{-x_0}^{x_0} f_0(x) \log \frac{\int_K \frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right) dP(\tau,\sigma)}{\int_K \frac{1}{\sigma}\psi\left(\frac{x-\tau-q_P}{\sigma}\right) dP(\tau,\sigma)} dx + \int_{|x|>x_0} f_0(x) \log \frac{f_0(x)}{f(x)} dx \\
<&\epsilon.
\end{aligned}
$$

This completes the proof. ■

Lemma 3.1.3 *Suppose all the assumptions in Lemma 3.1.2 hold, then* $\tilde{\Pi}(\mathcal{A} \cap \mathcal{E} \cap \mathcal{G}) > 0$.

Proof The family of functions $\left\{\frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right) : x \in [-x_0, x_0]\right\}$, viewed as a set of functions of (τ, σ) in K, is uniformly equicontinuous. By the Arzela-Ascoli theorem [34], there exist finitely many points $x_1, x_2, \cdots, x_m$ such that for any $x \in [-x_0, x_0]$, there exists an $i \in \{1, 2, \cdots, m\}$ with

$$
\sup_{(\tau,\sigma)\in K} \left| \frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right) - \frac{1}{\sigma}\psi\left(\frac{x_i-\tau}{\sigma}\right) \right| < c\kappa. \tag{3.10}
$$

Let

$$
\begin{aligned}
\mathcal{S} := \Bigg\{ P : \Bigg| &\int_K \frac{1}{\sigma}\psi\left(\frac{x_i-\tau}{\sigma}\right) dP_0^{\sigma_1}(\tau,\sigma) - \int_K \frac{1}{\sigma}\psi\left(\frac{x_i-\tau}{\sigma}\right) dP(\tau,\sigma) \Bigg| \\
&< c(1-\lambda)\kappa;\ i = 1, \cdots, m \Bigg\}.
\end{aligned}
$$

Note that $\mathcal{S}$ is a weak neighbourhood of $P_0^{\sigma_1}$. Since $\sigma_1 < h_0$ and by assumption $(C2)$ we have $\tilde{\Pi}(\mathcal{S}) > 0$.

If $P \in \mathcal{S} \cap \mathcal{E}$, then for any i,

$$\begin{aligned}&\left| \int_K \frac{1}{\sigma}\psi\left(\frac{x_i-\tau}{\sigma}\right) dP_0^{\sigma_1 *}(\tau,\sigma) - \frac{P(K)}{P_0^{\sigma_1}(K)} \int_K \frac{1}{\sigma}\psi\left(\frac{x_i-\tau}{\sigma}\right) dP^*(\tau,\sigma) \right| \\ <&\frac{c(1-\lambda)\kappa}{P_0^{\sigma_1}(K)} < c\kappa.\end{aligned}$$

Therefore, a triangulation argument leads to

$$\begin{aligned}&\left| \int_K \frac{1}{\sigma}\psi\left(\frac{x_i-\tau}{\sigma}\right) dP_0^{\sigma_1 *}(\tau,\sigma) - \int_K \frac{1}{\sigma}\psi\left(\frac{x_i-\tau}{\sigma}\right) dP^*(\tau,\sigma) \right| \\ <&ck + \left| \frac{P(K)}{P_0^{\sigma_1}(K)} - 1 \right| < (c+1)\kappa.\end{aligned}$$

So for any $x \in [-x_0, x_0]$, choosing an appropriate x_i from (3.10) and using a triangulation argument, we get

$$\log\left(\frac{\int\limits_K \frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right) dP_0^{\sigma_1 *}(\tau,\sigma)}{\int\limits_K \frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right) dP^*(\tau,\sigma)} \right) \leq \left| \frac{\int_K \frac{1}{\sigma}\psi\left(\frac{x_i-\tau}{\sigma}\right) dP_0^{\sigma_1 *}(\tau,\sigma)}{\int_K \frac{1}{\sigma}\psi\left(\frac{x_i-\tau}{\sigma}\right) dP^*(\tau,\sigma)} - 1 \right| < 3\kappa + \frac{\kappa}{c}.$$

Therefore, $\mathcal{S} \cap \mathcal{E} \subseteq \mathcal{G} \cap \mathcal{E}$. Since $\mathcal{S}$ and $\mathcal{E}$ are independent, $\tilde{\Pi}(\mathcal{S} \cap \mathcal{E}) = \tilde{\Pi}(\mathcal{S})\tilde{\Pi}(\mathcal{E}) > 0$. So $\tilde{\Pi}(\mathcal{G}) > \tilde{\Pi}(\mathcal{G} \cap \mathcal{E}) > 0$. Since $\mathcal{G}$ and $\mathcal{A} \cap \mathcal{E}$ are independent, $\tilde{\Pi}(\mathcal{A} \cap \mathcal{E} \cap \mathcal{G}) > 0$. ■

Lemma 3.1.4 *Suppose all the assumptions in Lemma 3.1.2 hold, then* $\tilde{\Pi}(\mathcal{A} \cap \mathcal{E} \cap \mathcal{G} \cap \mathcal{J}) > 0$.

Proof First by Lemma 3.1.1, we can pick a large constant $M_1 > 0$, and define set $\mathcal{S}_1 := \{P : |q_P| \leq M_1\}$ such that $\tilde{\Pi}(\mathcal{S}_1) > 1 - \tilde{\Pi}(\mathcal{A} \cap \mathcal{E} \cap \mathcal{G})$. Thus $\tilde{\Pi}(\mathcal{A} \cap \mathcal{E} \cap \mathcal{G} \cap \mathcal{S}_1) > 0$.

For any $P \in \mathcal{A} \cap \mathcal{E} \cap \mathcal{G} \cap \mathcal{S}_1$, q_P is bounded. Then by the uniformly continuity of the quantile function on the compact set $[-M_1, M_1]$, there exists a constant δ_1, such that if

$$\left| \int\limits_{-\infty}^{a} \int \frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right) dP(\tau,\sigma)dx - \int\limits_{-\infty}^{b} \int \frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right) dP(\tau,\sigma)dx \right| < \delta_1,$$

then $|a-b| < t_0$, where t_0 is the constant in the definition of $\mathcal{J}$.

Therefore, it suffices to show that by choosing proper x_0, σ_1, λ and κ, for any $P \in \mathcal{A} \cap \mathcal{E} \cap \mathcal{G} \cap \mathcal{S}_1$, we can find such a bound δ_1 which can be arbitrarily small for $a = q_P$ and

$b = 0$. Note that although the definition of t_0 depends on x_0, for large x_0, increasing x_0 will not decrease t_0 due to uniform continuity of the logistic density function. To simply our argument, we further consider $P \in \mathcal{A} \cap \mathcal{E} \cap \mathcal{S} \cap \mathcal{S}_1$, where $\mathcal{S}$ is defined in Lemma 3.1.3. As shown in Lemma 3.1.3, $\mathcal{A} \cap \mathcal{E} \cap \mathcal{S} \cap \mathcal{S}_1$ is a subset of $\mathcal{A} \cap \mathcal{E} \cap \mathcal{G} \cap \mathcal{S}_1$.

$$
\begin{aligned}
&\left| p - \int_{-\infty}^{0} \int \frac{1}{\sigma} \psi\left(\frac{x-\tau}{\sigma}\right) dP(\tau,\sigma)dx \right| \\
\leq &\left| p - \int_{-\infty}^{0} \int \frac{1}{\sigma} \psi\left(\frac{x-\tau}{\sigma}\right) dP_0^{\sigma_1}(\tau,\sigma)dx \right| \\
&+ \left| \int_{-x_0}^{0} \int_{K} \frac{1}{\sigma} \psi\left(\frac{x-\tau}{\sigma}\right) dP_0^{\sigma_1}(\tau,\sigma)dx - \int_{-x_0}^{0} \int_{K} \frac{1}{\sigma} \psi\left(\frac{x-\tau}{\sigma}\right) dP(\tau,\sigma)dx \right| \qquad (3.11) \\
&+ \left| \int_{-x_0}^{0} \int_{K^c} \frac{1}{\sigma} \psi\left(\frac{x-\tau}{\sigma}\right) dP_0^{\sigma_1}(\tau,\sigma)dx - \int_{-x_0}^{0} \int_{K^c} \frac{1}{\sigma} \psi\left(\frac{x-\tau}{\sigma}\right) dP(\tau,\sigma)dx \right| \\
&+ \left| \int_{-\infty}^{-x_0} \int \frac{1}{\sigma} \psi\left(\frac{x-\tau}{\sigma}\right) dP_0^{\sigma_1}(\tau,\sigma)dx - \int_{-\infty}^{-x_0} \int \frac{1}{\sigma} \psi\left(\frac{x-\tau}{\sigma}\right) dP(\tau,\sigma)dx \right|.
\end{aligned}
$$

We bound the four terms separately. For the first term, by Fubini's theorem and integration by parts, we have

$$
\int_{-\infty}^{0} \int \frac{1}{\sigma} \psi\left(\frac{x-\tau}{\sigma}\right) dP_0^{\sigma_1}(\tau,\sigma)dx = \int \frac{1}{\sigma_1} \psi\left(-\frac{\tau}{\sigma_1}\right) F_0(\tau)d\tau,
$$

where $F_0(x)$ is the cumulative distribution function of $f_0(x)$. Thus

$$
\int \frac{1}{\sigma_1} \psi\left(-\frac{\tau}{\sigma_1}\right) F_0(\tau)d\tau \ \to F_0(0) = p \quad \text{as} \ \ \sigma_1 \to 0.
$$

By choosing σ_1 properly, the first term can be made arbitrarily small.

For the second term,

$$
\begin{aligned}
&\left| \int_{-x_0}^{0} \int_{K} \frac{1}{\sigma} \psi\left(\frac{x-\tau}{\sigma}\right) dP_0^{\sigma_1}(\tau,\sigma)dx - \int_{-x_0}^{0} \int_{K} \frac{1}{\sigma} \psi\left(\frac{x-\tau}{\sigma}\right) dP(\tau,\sigma)dx \right| \\
\leq &\int_{-x_0}^{0} \int_{K} \frac{1}{\sigma} \psi\left(\frac{x-\tau}{\sigma}\right) dP(\tau,\sigma) \left| 1 - \frac{\int_{K} \frac{1}{\sigma} \psi\left(\frac{x-\tau}{\sigma}\right) dP_0^{\sigma_1}(\tau,\sigma)}{\int_{K} \frac{1}{\sigma} \psi\left(\frac{x-\tau}{\sigma}\right) dP(\tau,\sigma)} \right| dx.
\end{aligned}
$$

Since $P \in \mathcal{S}$, as shown in Lemma 3.1.3, $\left|1 - \frac{\int_K \frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right)dP_0^{\sigma_1}(\tau,\sigma)}{\int_K \frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right)dP(\tau,\sigma)}\right|$ can be arbitrarily small be choosing κ properly. Thus the bound of the second term can be arbitrarily small.

For the third term, since

$$\begin{aligned}&\left|\int_{-x_0}^{0}\int_{K^c}\frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right)dP_0^{\sigma_1}(\tau,\sigma)dx\right|\\ =&\int_{K^c}\left|\frac{1}{1+\frac{1-p}{p}\exp\left(\frac{\tau}{\sigma}\right)}-\frac{1}{1+\frac{1-p}{p}\exp\left(\frac{\tau+x_0}{\sigma}\right)}\right|dP_0^{\sigma_1}(\tau,\sigma)\le 2P_0^{\sigma_1}(K^c),\end{aligned}$$

and similarly

$$\begin{aligned}&\left|\int_{-x_0}^{0}\int_{K^c}\frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right)dP_0^{\sigma_1}(\tau,\sigma)dx\right|\\ =&\int_{K^c}\left|\frac{1}{1+\frac{1-p}{p}\exp\left(\frac{\tau}{\sigma}\right)}-\frac{1}{1+\frac{1-p}{p}\exp\left(\frac{\tau+x_0}{\sigma}\right)}\right|dP(\tau,\sigma)\le 2P(K^c).\end{aligned}$$

Since $P \in \mathcal{E}$, by the definition we have $P_0^{\sigma_1}(K^c) < \lambda$ and

$$P(K^c) < 1 - (1-\kappa)P_0^{\sigma_1}(K) = P_0^{\sigma_1}(K^c) + \kappa P_0^{\sigma_1}(K) < \lambda + \kappa.$$

Thus, the third term has an arbitrarily small bound when κ and λ are chosen appropriately.

For the last term, by Fubini's theorem and integration by parts, we have

$$\int_{-\infty}^{-x_0}\int\frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right)dP_0^{\sigma_1}(\tau,\sigma)dx=\int\frac{1}{\sigma_1}\psi\left(\frac{-x_0-\tau}{\sigma_1}\right)F_0(\tau)d\tau,$$

and

$$\int\frac{1}{\sigma_1}\psi\left(\frac{-x_0-\tau}{\sigma_1}\right)F_0(\tau)d\tau\to F_0(-x_0)\ \text{ as }\ \sigma_1\to 0.$$

Thus, $\left|\int_{-\infty}^{-x_0}\int\frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right)dP_0^{\sigma_1}(\tau,\sigma)dx\right|$ can be arbitrarily small by choosing large x_0 and small σ_1. On the other hand,

$$\begin{aligned}\left|\int_{-\infty}^{-x_0}\int\frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right)dP(\tau,\sigma)dx\right|&\le\left|\int_{-\infty}^{-x_0}f_0(x)\left(1-\frac{f(x)}{f_0(x)}\right)dx\right|+\int_{-\infty}^{-x_0}f_0(x)dx\\ &\le\int_{-\infty}^{-x_0}f_0(x)\log\frac{f_0(x)}{f(x)}dx+F_0(-x_0).\end{aligned}$$

Since $P \in \mathcal{A}$, $\int_{-\infty}^{-x_0} f_0(x) \log \frac{f_0(x)}{f(x)} dx$ and $F_0(-x_0)$ can be arbitrarily small by picking a large x_0. This concludes the proof. ■

With Lemma 3.1.2, we are ready to establish the Kullback-Leibler property for Π^*.

Theorem 3.1.1 *Suppose Π^* is a location-scale mixture prior obtained by defining the random probability density $f(x) = (\psi * P)(x - q_P)$ where $P \sim DP(\alpha, G)$. Let $\tilde{\Pi}$ denote $DP(\alpha, G)$.*

Assume that assumptions $(C1)$ and $(C2)$ in Lemma 3.1.2 hold, and the true density $f_0 \in \mathcal{F}_p$ satisfies

$(C3)$ $\int f_0 \log f_0(x) dx < \infty$;

$(C4)$ *there exists $0 < \eta < 1$ such that $\int |x|^\eta f_0(x) dx < \infty$.*

Further assume that there exist $\sigma_0 > 0$, $0 < \xi < \eta$, $\gamma > \xi$ and $b_1, b_2 > 0$ such that for large $x > 0$,

$(C5)$ $\max\left\{ G\left(\left[x - \frac{\sigma_0}{2} x^\eta, \infty \right) \times [\sigma_0, \infty) \right),\ G\left([0, \infty) \times (x^{1-\frac{\eta}{2}}, \infty) \right) \right\} \geq b_1 x^{-\xi}$;

$(C6)$ $\max\left\{ G\left(\left(-\infty, -x + \frac{\sigma_0}{2} x^\eta \right] \times [\sigma_0, \infty) \right),\ G\left((\infty, 0] \times (x^{1-\frac{\eta}{2}}, \infty) \right) \right\} \geq b_1 x^{-\xi}$;

$(C7)$ $G\left((-\infty, x) \times (0, x) \right) > 1 - b_2 x^{-\gamma}$;

$(C8)$ $G\left((-x, \infty) \times (0, x) \right) > 1 - b_2 x^{-\gamma}$.

Then, Π^* possesses the Kullback-Leibler property.

Proof Define

$$K_{x,P} := \left\{ (\tau, \sigma) \in \mathbb{R} \times \mathbb{R}^+ :\ \frac{1}{\sigma} \psi\left(\frac{x - \tau - q_P}{\sigma} \right) \geq \exp\left(-|x|^\eta \right) \right\}.$$

Then

$$\begin{aligned}
&\int_{|x|>x_0} f_0(x)\log\frac{f_0(x)}{f(x)}dx \\
\le &\int_{|x|>x_0} f_0(x)\log\frac{f_0(x)}{\int_{K_{x,P}}\frac{1}{\sigma}\psi\left(\frac{x-\tau-q_P}{\sigma}\right)dP(\tau,\sigma)}dx \\
\le &\int_{|x|>x_0} f_0(x)\log\frac{f_0(x)}{\exp(-|x|^\eta)P(K_{x,P})}dx
\end{aligned}$$

If we can show that for any $\epsilon>0$ there exist $x_0>0$ and a set $\mathcal{A}$ with $\tilde{\Pi}(A)>1-\epsilon$ such that $P\in\mathcal{A}$ implies $P(K_{x,P})>c_1\exp(-c_2|x|^\eta)$ for all $|x|>x_0$ with some fixed constants $c_1,c_2>0$. Then if $p\in\mathcal{A}$,

$$\begin{aligned}
&\int_{|x|>x_0} f_0(x)\log\frac{f_0(x)}{f(x)}dx \\
\le &\int_{|x|>x_0} f_0(x)\log\frac{f_0(x)}{\exp(-|x|^\eta)c_1\exp(-c_2|x|^\eta)}dx \\
\le &\int_{|x|>x_0} f_0(x)\left(\log f_0(x)+|x|^\eta+|\log c_1|+c_2|x|^\eta\right)dx.
\end{aligned}$$

By assumptions $(C3)$ and $(C4)$, $\int_{|x|>x_0} f_0(x)\left(\log f_0(x)+|x|^\eta+|\log c_1|+c_2|x|^\eta\right)dx$ can be made arbitrarily small, then by Lemma 3.1.2, Π^* possesses the Kullback-Leibler property.

Therefore, it suffices to state and prove Lemma 3.1.5 in the following. ■

Lemma 3.1.5 *Assume all the assumptions in Theorem 3.1.1 hold. Then for any $\epsilon>0$, there exists a constant $x_0>0$ and a set $\mathcal{A}$ with $\tilde{\Pi}(\mathcal{A})>1-\epsilon$ such that $P\in\mathcal{A}$ implies $P(K_{x,P})\ge c_1\exp(-c_2|x|^\eta)$ for all $|x|>x_0$, where c_1 and c_2 are constants.*

Proof An equivalent definition of $K_{x,P}$ is given by

$$\left\{(\tau,\sigma):\ x-q_P-b_x(\sigma)\le\tau\le x-q_P+b_x(\sigma), 0<\sigma\le\frac{1}{4}\exp(|x|^\eta)\right\},$$

where $b_x(\sigma)=2\sigma\log\left(\frac{1}{2\sqrt{\sigma}}\exp\left(\frac{|x|^\eta}{2}\right)+\sqrt{\frac{1}{4\sigma}\exp\left(|x|^\eta\right)-1}\right)$. After some simple algebra it can be seen that as $\sigma\to 0$ or $\sigma\to\frac{1}{4}\exp(|x|^\eta)$, $b_x(\sigma)\to 0$ and the maximum of $b_x(\sigma)$ is

attained at $\sigma_{m,x} = \frac{z_0^2-1}{4z_0^2}\exp(|x|^\eta)$, where z_0 is the solution to the equation $\exp(z) = \frac{z+1}{z-1}$ for $z > 0$.

It is hard to directly work with $K_{x,P}$ which is indexed by both x and P, we define a subset of $K_{x,P}$ in the following way. Pick a large $M > 0$ such that $b_x(\sigma_{m,x}) > |x|^{\frac{\eta}{2}}$ for all $|x| > M$. And let $\sigma_{l,x} < \sigma_{r,x}$ be the two solutions of $b_x(\sigma) = |x|^{\frac{\eta}{2}}$. For $x > M$, let

$$\tilde{K}_x := \left\{(\tau,\sigma) :\ x + |x|^{\frac{\eta}{2}} - b_x(\sigma) \le \tau \le x - |x|^{\frac{\eta}{2}} + b_x(\sigma), \sigma_{l,x} \le \sigma \le \sigma_{r,x}\right\}.$$

We will show in Lemma 3.1.6 that there exist a constant $x_1 > 0$ and a set $\mathcal{B}$ with $\tilde{\Pi}(\mathcal{B}) > 1 - \frac{\epsilon}{2}$ such that $P \in \mathcal{B}$ implies $P(\tilde{K}_x) \ge c_1\exp(-c_2|x|^\eta)$ for all $|x| > x_1$, where c_1 and c_2 are constants. On the other hand, by Lemma 3.1.1, there exist a constant $x_2 > 0$ and a set $\mathcal{C} := \{P :\ |q_P| \le x_2\}$ such that $\tilde{\Pi}(\mathcal{C}) \ge 1 - \frac{\epsilon}{2}$. Therefore, we have

$$\tilde{\Pi}(\mathcal{B}\cap\mathcal{C}) = \tilde{\Pi}(\mathcal{B}) + \tilde{\Pi}(\mathcal{C}) - \tilde{\Pi}(\mathcal{B}\cup\mathcal{C}) > 1 - \frac{\epsilon}{2} + 1 - \frac{\epsilon}{2} - 1 = 1 - \epsilon.$$

For $P \in \mathcal{B}\cap\mathcal{C}$, if $|x| > x_2^{2/\eta}$, by the definitions of $K_{x,P}$ and $\tilde{K}_x$, we have $\tilde{K}_x \subseteq K_{x,P}$, thus $P(K_{x,P}) \ge P(\tilde{K}_x) \ge c_1\exp(-c_2|x|^\eta)$ for all $|x| > \max\{x_1, x_2^{2/\eta}\}$. This completes the proof. ■

Lemma 3.1.6 *Assume all the assumptions in Theorem 3.1.1 hold. Then for any $\epsilon > 0$, there exists $x_0 > 0$ and a set $\mathcal{A}$ with $\tilde{\Pi}(\mathcal{A}) > 1 - \epsilon$ such that $P \in \mathcal{A}$ implies $P(\tilde{K}_x) \ge c_1\exp(-c_2|x|^\eta)$ for all $|x| > x_0$, where c_1 and c_2 are constants.*

Proof Using the approach in [102], we first prove the result for $x \ge 0$, and the case where $x \le 0$ can be proved using the same argument. Pick $t_0 > 0$ such that $x + x^{\frac{\eta}{2}} - b_x(\sigma_0)$ as a function of x is monotonically increasing for $x \in \{x : x > t_0, \sigma_0 \le \frac{1}{4}\exp(x^\eta)\}$. The existence of such t_0 and σ_0 can be seen from the fact that, for large x,

$$b_x(\sigma_0) = \sigma_0 x^\eta + 2\sigma_0\log\left(\frac{1}{2\sqrt{\sigma_0}} + \sqrt{\frac{1}{4\sigma_0} - \exp(-x^\eta)}\right) \in \left(\frac{\sigma_0}{2}x^\eta, 2\sigma_0 x^\eta\right). \tag{3.12}$$

Define for $x \geq 0$,

$$A_x := \begin{cases} \mathbb{R} \times \mathbb{R}^+, & x \leq t_0, \\ \left\{ [x + x^{\frac{\eta}{2}} - b_x(\sigma_0), \infty) \times [\sigma_0, \infty) \right\} \cup \left\{ [0, \infty) \times (x^{1-\frac{\eta}{2}}, \infty) \right\}, & x > t_0. \end{cases}$$

and

$$B_x := \begin{cases} \mathbb{R} \times \mathbb{R}^+, & x = 0, \\ \{(x, \infty) \times (\sigma_0, \infty)\} \cup \{[0, \infty) \times (\sigma_{m,x}, \infty)\}, & x > 0. \end{cases}$$

Then let us investigate some properties of A_x and B_x. Clearly

$$B_x \subseteq A_x \text{ for all } x \geq 0. \tag{3.13}$$

Also by the monotonicity of $x + x^{\frac{\eta}{2}} - b_x(\sigma_0)$, $x^{1-\frac{\eta}{2}}$, x and $\sigma_{m,x}$, we have that

$$0 < x < y \quad \text{implies} \quad A_x^c \subseteq A_y^c \text{ and } B_x^c \subseteq B_y^c. \tag{3.14}$$

Besides from (3.12) and $0 < \eta < 1$, $\lim_{x \to \infty} x + x^{\frac{\eta}{2}} - b_x(\sigma_0) = \infty$. So

$$A_0^c = B_0^c = \varnothing \text{ and } A_x^c, B_x^c \uparrow \mathbb{R} \times \mathbb{R}^+ \text{ as } x \uparrow \infty. \tag{3.15}$$

Moreover, A_x and B_x are closely connected with $\tilde{K}_x$. Let $A_x \backslash B_x$ denote the set difference, then

$$A_x \backslash B_x = \left\{ [x + x^{\frac{\eta}{2}} - b_x(\sigma_0), x] \times [\sigma_0, \sigma_{m,x}] \right\} \cup \left\{ [0, x] \times (x^{1-\frac{\eta}{2}}, \sigma_{m,x}] \right\}.$$

From (3.12) for large x, $b_x(\sigma_0) > x^{\frac{\eta}{2}}$, thus $\sigma_0 \in [\sigma_{l,x}, \sigma_{r,x}]$ for large x. Similarly, for large x, $b_x\left(x^{1-\frac{\eta}{2}}\right) > x^{\frac{\eta}{2}}$, so $x^{1-\frac{\eta}{2}} \in [\sigma_{l,x}, \sigma_{r,x}]$ for large x. Also for large x, $b_x(\sigma) - |x|^{\frac{\eta}{2}} > 0$ for $\sigma \in [\sigma_0, \sigma_{m,x}]$ or $\sigma \in (x^{1-\frac{\eta}{2}}, \sigma_{m,x}]$, thus there exists a constant $x_1 > 0$ such that

$$A_x \backslash B_x \subset \tilde{K}_x \text{ for } x > \max\{x_1, M\}. \tag{3.16}$$

Equations (3.13), (3.14) and (3.15) ensure that $F_A(x) = P(A_x^c)$ and $F_B(x) = P(B_x^c)$ are random probability functions on $[0, \infty)$. Moreover, F_A and F_B follows $DP(\alpha, G_A)$

and $DP(\alpha, G_B)$ respectively, where $G_A([0,x]) = G(A_x^c)$ and $G_B([0,x]) = G(B_x^c)$. Then it follows from [25],

$$Pr\left(\limsup_{x\to\infty} \frac{P(B_x)}{h_1(G(B_x)} = 0\right) = 1,$$

and

$$Pr\left(\liminf_{x\to\infty} \frac{P(A_x)}{h_2(G(A_x)} = \infty\right) = 1,$$

where $h_1(t) = \exp\left(-\frac{1}{t|\log t|^2}\right)$ and $h_2(t) = \exp\left(-\frac{2\log|\log t|}{t}\right)$ for small $t > 0$. Let $\mathcal{M}$ be the space of all probability measures on $(\mathbb{R}, \mathcal{B}_\mathbb{R})$. Let $\mathcal{B}_\mathcal{M}$ be the smallest σ-algebra generated by all weakly open sets. Then $\frac{P(B_x)}{h_1(G(B_x)}$ and $\frac{P(A_x)}{h_2(G(A_x)}$ are both $\mathcal{B}_\mathcal{M}$-measurable function. Then by applying Egoroff's theorem [34] to the two sequences of functions indexed by x, for any $\epsilon > 0$, there exist $x_2 > 0$ and a set $\mathcal{A}$ with $\tilde{\Pi}(\mathcal{A}) > 1 - \frac{\epsilon}{2}$, such that $P \in \mathcal{A}$ implies $P(B_x) < h_1(G(B_x))$ and $P(A_x) > h_2(G(A_x))$ for any $x > x_2$.

Thus if $P \in \mathcal{A}$, for $x > \max\{M, x_1, x_2\}$,

$$\begin{aligned}
&P(\tilde{K}_x) \geq P(A_x \backslash B_x) \geq h_2(G(A_x)) - h_1(G(B_x)) \\
=&\exp\left(-\frac{2\log|\log G(A_x)|}{G(A_x)}\right) - \exp\left(-\frac{1}{G(B_x)(\log G(B_x))^2}\right) \\
=&\exp\left(-\frac{2\log|\log G(A_x)|}{G(A_x)}\right)\left(1 - \exp\left(\frac{2\log|\log G(A_x)|}{G(A_x)} - \frac{1}{G(B_x)(\log G(B_x))^2}\right)\right).
\end{aligned} \tag{3.17}$$

From (3.12), $x - \frac{\sigma_0}{2}x^\eta > x + x^{\frac{\eta}{2}} - b_x(\sigma_0)$ for large $x > 0$, and from assumption $(C1)$, we have $G(A_x) \geq b_1 x^{-\xi}$ for large $x > 0$. By assumption $(C3)$,

$$G(B_x) \leq G((\{(-\infty, x) \times (0, x)\}^c)) \leq b_2 x^{-\gamma} \text{ for large } x > 0.$$

Therefore, for large $x > 0$, since $\xi < \eta$ and $\gamma > \xi$, we have

$$\begin{aligned}
&\exp\left(-\frac{2\log|\log G(A_x)|}{G(A_x)}\right) = \exp\left(-\frac{2\log(-\log G(A_x))}{G(A_x)}\right) \\
\geq&\exp\left(-\frac{2\log(\xi\log x - \log b_1)}{b_1 x^{-\xi}}\right) \geq \exp\left(-\frac{2}{b_1}x^\eta\right),
\end{aligned} \tag{3.18}$$

and for x large enough

$$\begin{aligned} &1-\exp\left(\frac{2\log|\log G(A_x)|}{G(A_x)}-\frac{1}{G(B_x)(\log G(B_x))^2}\right)\\ \geq &1-\exp\left(\frac{2\log(\xi\log x-\log b_1)}{b_1x^{-\xi}}-\frac{1}{b_2x^{-\gamma}(\gamma\log x-\log b_2)^2}\right)\geq\frac{1}{2}. \end{aligned} \tag{3.19}$$

Plugging (3.18) and (3.19) into (3.17), we conclude that there exists $x_0>\max\{M,x_1,x_2\}$ such that if $x>x_0$ then $P(\tilde{K}_x)\geq\frac{1}{2}\exp\left(-\frac{2}{b_1}x^\eta\right)$.

■

3.2 Posterior consistency for quantile regression

Consider a simple case where the true model is $Y=\beta_0^*+\beta_1^*x+\epsilon$ with true probability density f_0 of ϵ has its p-th quantile as 0.

Follow the notations in [4, 102], given any pdf f, define $f_{\beta_0,\beta_1,i}(y):=f(y-\beta_0-\beta_1x_i)$ and $f_{0i}(y)=f_0(y-\beta_0^*-\beta_1^*x_i)$. For any two probability densities f and g, let $K(f,g):=\int f(x)\log\frac{f(x)}{g(x)}dx$ and $V(f,g):=\int f(x)\left(\log_+\frac{f(x)}{g(x)}\right)^2dx$, where $\log_+(x):=\max\{0,\log x\}$. Then define $K_i(f,\beta_0,\beta_1):=K(f_{0i},f_{\beta_0,\beta_1,i})$ and $V_i(f,\beta_0,\beta_1)=V(f_{0i},f_{\beta_0,\beta_1,i})$.

Now we define the exponentially consistent sequence of test functions which plays a key role in the Schwartz's theorem [4, 93].

Definition 3.2.1 *A test function is a non-negative measurable function bounded by 1.*

Definition 3.2.2 *Let $\mathcal{W}\subset\mathcal{F}_p\times\mathbb{R}\times\mathbb{R}$. A sequence of test functions $\mathbf{\Phi}_n(Y_1,\cdots,Y_n)$ is said to be exponentially consistent for testing*

$$H_0:(f,\beta_0,\beta_1)=(f_0,\beta_0^*,\beta_1^*)\ \ \textit{against}\ \ H_1:(f,\beta_0,\beta_1)\in\mathcal{W},$$

if there exist constants $c,c_1,c_2>0$ such that

(a) $E_{\prod_{i=1}^n f_{0i}}\mathbf{\Phi}_n\leq c_1\exp(-nc)$;

(b) $\inf_{(f,\beta_0,\beta_1)\in\mathcal{W}} E_{\prod_{i=1}^{n} f_{\beta_0,\beta_1,i}} \mathbf{\Phi}_n \geq 1 - c_2 \exp(-nc)$.

Our main tool to show the posterior consistency is a modification to the result presented in [4] which is a variant of Schwartz's theorem [93] .

Theorem 3.2.1 *Let* Π *be a prior over* $\mathcal{F}_p \times \mathbb{R} \times \mathbb{R}$. *If*

(1) there exists $\epsilon_0 > 0$ *such that*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{x_i < -\epsilon_0\} > 0, \qquad \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{x_i > \epsilon_0\} > 0,$$

(2) for some $L > 0$, $|x_i| < L$ *for all* i,

(3) and for all $\delta > 0$,

$$\Pi\left\{(f,\beta_0,\beta_1) :\ K_i(f,\beta_0,\beta_1) < \delta \text{ for all } i,\ \sum_{i=1}^{\infty} \frac{V_i(f,\beta_0,\beta_1)}{i^2} < \infty\right\} > 0,$$

then Π *has weak posterior consistency at* $(f_0, \beta_0^*, \beta_0^*)$.

Proof Theorem 2.1 in [4] adapts Schwartz's theorem to the regression case with the error density in the space of symmetric densities. Although our $\mathcal{F}_p$ is different and contains all the densities with the p-th quantile equal to 0, the proof still carries over with no modification required. But to apply this variant of Schwartz's theorem, we need to show the existence of an exponentially consistent sequences of test functions. For all the proofs in Section 3 of [4], only Lemma 3.2 requires the symmetry of the densities. So it suffices to provide a similar lemma for the densities subject to the quantile constraint. We conclude the proof by stating and proving Lemma 3.2.1. Note that our Lemma 3.2.1 can be used in the same fashion as Lemma 3.2 of [4] in proving Propositions 3.1 and 3.2 in [4]. ■

For a density g and $\theta \in \mathbb{R}$, let $g_\theta(y) := g(y - \theta)$.

Lemma 3.2.1 *Let g_0 be a density on $\mathbb{R}$ with the p-th quantile equal to 0. Let η be such that, if $\inf_{|y|<\eta} g_0(y) = c > 0$,*

(i) for any $\Delta > 0$, there exists a set $B_\Delta \subseteq \mathbb{R}$ such that

$$P_{g_0}(B_\Delta) \leq 1 - p - c\min\{\Delta, \eta\},$$

and, for any density g with $Q_g(p) = 0$,

$$P_{g_\theta}(B_\Delta) \geq 1 - p, \quad \textit{for all } \theta \geq \Delta;$$

(ii) for any $\Delta < 0$, there exists a set $\tilde{B}_\Delta \subseteq \mathbb{R}$ such that

$$P_{g_0}(\tilde{B}_\Delta) \leq p - c\min\{-\Delta, \eta\},$$

and, for any density g with $Q_g(p) = 0$,

$$P_{g_\theta}(\tilde{B}_\Delta) \geq p, \quad \textit{for all } \theta \geq \Delta.$$

Proof (i) Take $B_\Delta = (\Delta, \infty)$. Since $\theta \geq \Delta$,

$$P_{g_\theta}(B_\Delta) = \int_\Delta^\infty g_\theta(y)dy = \int_{\Delta-\theta}^\infty g(y)dy \geq 1 - p.$$

And

$$\begin{aligned} P_{g_0}(B_\Delta) &= \int_\Delta^\infty g_0(y)dy = 1 - p - \int_0^\Delta g(y)dy \\ &\leq 1 - p - \int_0^{\min\{\Delta,\eta\}} g(y)dy \leq 1 - p - c\min\{\Delta, \eta\}. \end{aligned}$$

(ii) Take $\tilde{B}_\Delta = (-\infty, \Delta)$. Similar argument completes the proof. ■

Now we have a variant of Swartz's theorem in our context. We then show the posterior consistency by establishing condition (3) of Theorem 3.2.1. Lemma 3.2.2, in the same

spirit as Lemma 6.1 in [4] and Lemma 5.1 in [102], provides a tool to link the Kullback-Leibler property of Π^* to the posterior consistency in the regression problem. Let $g_t(x) := g(x-t)$.

Lemma 3.2.2 *Fix a P, and take $g(x) = (\psi * P)(x - q_P)$. If*

1. $\int f_0(x)(\log f_0(x))^2 dx < \infty$;
2. *there exists $\eta > 0$ such that $\int |x|^{2\eta} f_0(x) dx < \infty$;*
3. $\sigma_P := \int \frac{1}{\sigma} dP(\tau, \sigma) < \infty$;
4. *there exist $M, b_1, b_2 > 0$, such that for all $|x| > M$,*

$$g(x) > b_1 \exp(-b_2 |x|^{\eta});$$

then,

(a) $\lim_{t\to 0} \int f_0(x) \log \frac{f_0(x)}{g_t(x)} dx = \int f_0(x) \log \frac{f_0(x)}{g(x)} dx$,
(b) $\lim_{t\to 0} \int f_0(x) \left(\log \frac{f_0(x)}{g_t(x)}\right)^2 dx = \int f_0(x) \left(\log \frac{f_0(x)}{g(x)}\right)^2 dx$.

Proof Although the normal distribution in [102] is now replaced by the logistic distribution, there is no significant modification required in the proof. We reproduce the proof here.

We first show that $g(x)$ is continuous. For the sequence of functions $\left\{\frac{1}{\sigma}\psi\left(\frac{x-\tau-q_P-t}{\sigma}\right)\right\}$ indexed by t, as $t \to 0$, $\frac{1}{\sigma}\psi\left(\frac{x-\tau-q_P-t}{\sigma}\right) \to \frac{1}{\sigma}\psi\left(\frac{x-\tau-q_P}{\sigma}\right)$. Assumption 3 guarantees that there is a P-integrable upper bound function $\frac{1}{\sigma}$. So we have $g(x-t) \to g(x)$ as $t \to 0$ by the dominated convergence theorem. Therefore, for each fixed x, as $t \to 0$,

$$\begin{aligned} f_0(x) \log \frac{f_0(x)}{g_t(x)} &\to f_0(x) \log \frac{f_0(x)}{g(x)}, \\ f_0(x) \left(\log \frac{f_0(x)}{g_t(x)}\right)^2 &\to f_0(x) \left(\log \frac{f_0(x)}{g(x)}\right)^2. \end{aligned} \tag{3.20}$$

Fix a $t_0 > 0$. For all $|x| > M + t_0$ and $|t| < t_0$, thus $|x - t| > M$, by Assumption 4,

$$g_t(x) = g(x - t) > b_1 \exp(-b_2|x - t|^\eta) \geq b_1 \exp(-b_2(|x| + |t_0|)^\eta).$$

Define $c_P := \inf\{g(x) : \ |x| \leq M + 2t_0\}$. Clearly $c_p > 0$. For any $|x| \leq M + t_0$ and $|t| < t_0$,

$$g_t(x) = g(x - t) \geq c_P.$$

Also for all x and t, we have

$$g_t(x) \leq \int \frac{1}{\sigma} dP(\tau, \sigma) = \sigma_P < \infty.$$

Let $h(x) := \max\{|\log b_1| + b_2(|x| + |t_0|)^\eta, |\log c_P|, |\log \sigma_P|\}$. Then for all x and $|t| < t_0$, $|\log g_t(x)| \leq h(x)$. For any x and $|t| < t_0$,

$$\left|\log \frac{f_0(x)}{g_t(x)}\right| \leq |\log f_0(x)| + |\log g_t(x)| \leq |\log f_0(x)| + h(x),$$

and

$$\begin{aligned}\left(\log \frac{f_0(x)}{g_t(x)}\right)^2 &= (\log f_0(x) - \log g_t(x))^2 \\ &\leq 2\log f_0(x))^2 + 2(\log g_t(x))^2 \leq 2\log f_0(x))^2 + 2(h(x))^2.\end{aligned}$$

Assumptions 1 and 2 guarantee

$$\int (|\log f_0(x)| + h(x)) f_0(x) dx < \infty, \ \int f_0(x)(2\log f_0(x))^2 + 2(h(x))^2) dx < \infty.$$

An application of the dominated convergence theorem [34] completes the proof. ■

Now we are ready to establish the posterior consistency for the quantile regression problem.

Theorem 3.2.2 *Suppose that Π^* is a prior over $\mathcal{F}_p$ specified by defining the random probability density $f(x) = (\psi * P)(x - q_P)$, where $P \sim DP(\alpha, G)$. Let $\tilde{\Pi}$ denote $DP(\alpha, G)$. Let π be a prior for (β_0, β_1). Let $\Pi := \Pi^* \times \pi$. Assume that*

1. *Assumptions (1) and (2) of Theorem 3.2.1 hold;*

2. *Assumptions* $(C5)$, $(C6)$, $(C7)$ *and* $(C8)$ *of Theorem 3.1.1 hold for some* $\sigma_0 > 0$, $0 < \xi < \eta < 1$, $\gamma > \xi$ *and* $b_1, b_2 > 0$*;*

3. *Assumptions* $(C1)$ *and* $(C2)$ *of Lemma 3.1.2 hold for some* $h_0 > 0$*;*

4. $\int f_0(x)(\log f_0(x))^2 dx < \infty$*;*

5. $\int |x|^{2\eta} f_0(x) dx < \infty$*;*

6. $\sigma_P = \int \frac{1}{\sigma} dP(\tau, \sigma) < \infty$ *almost surely.*

Then, Π *achieves weak posterior consistency at* $(f_0, \beta_0^*, \beta_1^*)$ *provided that* (β_0^*, β_1^*) *is in the support of* π*.*

Proof For $f(x) = \psi * P(x - q_P)$, by Lemma 3.1.5, for any $\epsilon > 0$, there exists a constant $x_0 > 0$ and a set $\mathcal{A}$ with $\tilde{\Pi}(\mathcal{A}) > 1 - \epsilon$ such that

$$\Pi^*(\{f \in \mathcal{F}_p : f(x) \geq c_1 \exp(-c_2|x|^\eta) \text{ for all} |x| > x_0\}) > 0. \tag{3.21}$$

And if $P \in \mathcal{A}$, by Assumptions 4 and 5 and Theorem 3.1.1, for any $\delta > 0$,

$$\Pi^*(\{f \in \mathcal{F}_p : K(f_0, f) < \delta,\ V(f_0, f) < \infty\}) > 0. \tag{3.22}$$

Also for any $P \in \mathcal{A}$, assumptions of Lemma 3.2.2 are satisfied. Put $\theta_i := \beta_0 - \beta_0^* + (\beta_1 - \beta_1^*)x_i$. Therefore, by Assumption 1, Assumption (2) of Theorem 3.2.1 and the assumption that (β_0^*, β_1^*) is in the support of π, with positive probability, $|\theta_i| < \delta$, then using Lemma 3.2.2,

$$\begin{aligned} K_i(f, \beta_0, \beta_1) &= \int f_0(y) \log \frac{f_0(y)}{f_{\theta_i}(y)} dx \\ &= \int f_0(y) \log \frac{f_0(y)}{f(y)} dy + \int f_0(y) \log \frac{f(y)}{f_{\theta_i}(y)} dy \\ &< K(f_0, f) + \delta, \end{aligned} \tag{3.23}$$

and

$$
\begin{aligned}
V_i(f, \beta_0, \beta_1) &= \int f_0(y) \left(\log \frac{f_0(y)}{f_{\theta_i}(y)} \right)^2 dy \\
&\leq 2 \int f_0(y) \left(\left(\log \frac{f_0(y)}{f(y)} \right)^2 + \left(\log \frac{f(y)}{f_{\theta_i}(y)} \right)^2 \right) dy \qquad (3.24) \\
&= 2V(f_0, f) + 2\delta.
\end{aligned}
$$

An application of Theorem 3.2.1 completes the proof.

■

Remark 3.1

Given independently identically distributed random variables X_i for $i = 1, \cdots, n$, Bayesian density estimation considers the model $p(x_i|f) = f(x_i)$ with $f \sim \Lambda$, where Λ is a prior over the space of pdfs and $p(x_i|f)$ is the pdf of X_i given f. A popular choice of Λ is the DPM model as in [29]. We can also consider the Bayesian density estimation subject to the quantile constraint, that is, we add a constraint $Q_{X_i}(p) = 0$ for $i = 1, \cdots, n$. Then we can model $p(x_i|f) = f(x_i)$ with $f \sim \Pi^*$. Then the Kullback-Leibler property established for Π^* in Section 3.1, combined with Schwartz's theorem [94] gives us the weak consistency of the density estimation problem subject to the quantile constraint. The proof that establishes weak consistency using Schwartz's theorem with Kullback-Leibler property can be found in many standard literatures, for example in Chapter 4.4 of [44].

Remark 3.2

Let $\mathcal{L}$ denote the space of all probability density functions. The weak consistency mentioned in Theorem 3.2.2 and Remark 3.1 are essentially with respect to the subspace topology on $\mathcal{F}_p$ induced from the weak topology on $\mathcal{L}$.

Remark 3.3

Note that in the DPML model (2.2) proposed in Chapter 2, we have

$$f(x) = \int \frac{1}{\sigma}\psi\left(\frac{x+\tau}{\sigma}\right) dP(\tau, \sigma),$$

where $P \sim DP(\alpha, G)$ and G is given in (2.4), while in this section the sign before τ is the opposite. However, by a simple calculation, we can rewrite the DPML model in Chapter 2 as $f(x) = \int \frac{1}{\sigma}\psi\left(\frac{x-\tau}{\sigma}\right) dP(\tau, \sigma)$ with $P \sim DP(\alpha, \tilde{G})$, where

$$\tilde{G} = \pi_\sigma(\tau) \cdot \text{Inv-Gamma}(\sigma|c, d),$$

with $\pi_\sigma(\tau)$ being the logistic distribution with its $(1-p)$-th quantile equal to 0 and the scale parameter being α. To guarantee the posterior consistency, the conditions on the base measure are all related to the tail behaviour. Because G and $\tilde{G}$ have the same tail behaviour, the theory derived in this section applies to our proposed model.

Remark 3.4

The assumptions on the tail of the base measure, $(C7)$ and $(C8)$ in Theorem 3.2.2, are stronger than the corresponding conditions in Theorem 3.3 in [102]. Basically, we require the base measure to have heavier tail. This requirement is because of the techniques used in the proof of Lemma 3.1.1. To bound the right tail of a beta distribution with the second parameter going to 0, we applied the Markov inequality. We believe a more delicate analysis of the incomplete beta function may relax these assumptions. However, the current assumptions are still general enough to incorporate our proposed model. The same argument in Remark 3.4 of [102] still holds. Thus the standard choice of normal-inverse-gamma distribution [29] satisfies Assumptions $(C7)$ and $(C8)$ in Theorem 3.2.2. And since our logistic-inverse-gamma distribution (2.4) has a heavier tail, these assumptions are automatically satisfied.

Remark 3.5

The assumption $(C1)$ can be satisfied for a wide range of true error densities as discussed in Remark 3 of [38]. For example, if $f_0(x)$ is continuous and monotone increasing on $(-\infty, a)$ and monotone decreasing on (b, ∞) for some $a < b$, and $\int f_0(x) \log f_0(x) dx < 0$, then by applying the dominated convergence theorem, Assumption $(C1)$ is satisfied.

Remark 3.6

The theory on posterior consistency developed in this chapter can be easily extended to the scenario where the logistic kernel is replaced by other kernels. For example, let ϕ denote the pdf of the standard normal distribution and consider the following prior over $\mathcal{F}_p$

$$f(x) = \int \frac{1}{\sigma} \phi\left(\frac{x - \mu - \tau_P}{\sigma}\right) dP(\mu, \sigma), \quad P \sim DP(\alpha, G), \tag{3.25}$$

where τ_P is defined in such a way that $\int_{-\infty}^{\tau_P} \int \frac{1}{\sigma} \phi\left(\frac{x-\mu-\tau_P}{\sigma}\right) dP(\tau, \sigma) dx = p$. The posterior consistency of (3.25) can be established by combining the results in this chapter and the results in [102], as long as we can show that Lemma 3.1.1 holds for τ_P. We state and prove such a lemma below.

Lemma 3.2.3 *Assume that there exist $\gamma > 0$ and $c > 0$ such that for large $x > 0$,*

$$G((-\infty, x) \times (0, x)) > 1 - cx^{-\gamma} \text{ and } G((-x, \infty) \times (0, x)) > 1 - cx^{-\gamma}.$$

Then there exists a constant $d > 0$ such that for any $t > 0$,

$$Pr(|\tau_P| \leq t) > 1 - \frac{d}{t^\gamma}.$$

Proof Pick $\nu \in (0, 1)$ such that $\Phi\left(\frac{1-\nu}{\nu}\right) > \max\{p, 1-p\}$, where Φ is the cumulative distribution function for the standard normal distribution. Define

$$a_\nu := \max\left\{\frac{p}{\Phi\left(\frac{1-\nu}{\nu}\right)}, 1 - p + \Phi\left(\frac{\nu - 1}{\nu}\right)\right\}.$$

Clearly $a_\nu < 1$. Also let $S_{t,\nu} := [-\nu t, \nu t] \times (0, \nu t)$. Then given any probability measure P over $\mathbb{R} \times \mathbb{R}^+$ if $P(S_{t,\nu}) \geq a_\nu$, we have

$$\begin{aligned}
&\int_{-\infty}^{-t} \int \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) dP(\mu,\sigma)dx = \int \Phi\left(\frac{-t-\mu}{\sigma}\right) dP(\mu,\sigma) \\
=&\int_{S_{t,\nu}} \Phi\left(\frac{-t-\mu}{\sigma}\right) dP(\mu,\sigma) + \int_{S_{t,\nu}^c} \Phi\left(\frac{-t-\mu}{\sigma}\right) dP(\mu,\sigma) \\
\leq&\Phi\left(\frac{\nu-1}{\nu}\right) P(S_{t,\nu}) + P\left(S_{t,\nu}^c\right) \\
\leq&\Phi\left(\frac{\nu-1}{\nu}\right) + 1 - a_\nu \leq p,
\end{aligned} \tag{3.26}$$

and

$$\begin{aligned}
&\int_{-\infty}^{t} \int \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) dP(\mu,\sigma)dx = \int \Phi\left(\frac{t-\mu}{\sigma}\right) dP(\mu,\sigma) \\
\geq&\int_{S_{t,\nu}} \Phi\left(\frac{t-\mu}{\sigma}\right) dP(\mu,\sigma) \geq \Phi\left(\frac{1-\nu}{\nu}\right) P(S_{t,\nu}) \geq p.
\end{aligned} \tag{3.27}$$

By equations (3.26) and (3.27), we have $|\tau_P| \leq t$. Therefore,

$$Pr(|\tau_P| \leq t) \geq Pr[P(S_{t,\nu}) \geq a_\nu] = Pr[1 - P(S_{t,\nu}) \leq 1 - a_\nu]. \tag{3.28}$$

Since $P \sim DP(\alpha, G)$, $1 - P(S_{t,\nu}) \sim$ Beta$(\alpha G(S_{t,\epsilon,\nu}^c), \alpha G(S_{t,\nu}))$. Therefore, applying the Markov inequality, we have

$$\begin{aligned}
&Pr[1 - P(S_{t,\nu}) \leq 1 - a_\nu] = 1 - Pr[1 - P(S_{t,\nu}) \geq 1 - a_\nu] \\
\geq&1 - \frac{E[1 - P(S_{t,\nu})]}{1 - a_\nu} = 1 - \frac{1 - G(S_{t,\nu})}{1 - a_\nu}.
\end{aligned} \tag{3.29}$$

Now by the assumption on the tail of G, $G(S_{t,\nu}) > 1 - 2ct^{-\gamma}$. Plug $G(S_{t,\nu})$ into equation (3.29), we get

$$Pr(1 - P(S_{t,\nu}) \leq 1 - a_\nu) >= 1 - \frac{2c}{(1 - a_\nu)t^\gamma}. \tag{3.30}$$

Note that a_ν does not depend on t. The result follows by combining equation (3.30) with equation (3.28). ■

Therefore, the posterior consistency of (3.25) can be established. And by Remark 3.4, we can specify G as the normal-inverse-gamma distribution, that is, $G(\mu, \sigma^2) = \mathrm{N}(\mu|0, \sigma^2) \cdot \text{Inv-Gamma}(\sigma^2|c, d)$. Now we have conjugacy between the base measure and the kernel, which can further simplify the MCMC algorithm. This motivates the application of DPM of normal distributions in Chapter 5.

4. Modelling heteroscedasticity

As shown in the simulation study in Section 2.4, when the data are heteroscedastic, DPM-based methods do not perform well if the heteroscedasticity is not explicitly modelled. In this chapter, we extend the DPML model proposed in Chapter 2 to heteroscedastic data.

4.1 The model

As in [53, 89], we assume the heteroscedastic linear regression model

$$Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + (\boldsymbol{x}_i^T \boldsymbol{\gamma})\epsilon_i, \text{ for } i = 1, \cdots, n, \tag{4.1}$$

where $Q_\epsilon(p) = 0$. Still we have $x_{i0} = 1$. Note that $\boldsymbol{\gamma}$ and parameters in the distribution of ϵ_i are unidentifiable, because given $(\boldsymbol{x}_i^T \boldsymbol{\gamma})\epsilon_i$ we can flip the sign of ϵ_i and find another $\tilde{\boldsymbol{\gamma}}$ such that $(\boldsymbol{x}_i^T \boldsymbol{\gamma}) = -(\boldsymbol{x}_i^T \tilde{\boldsymbol{\gamma}})$. Thus, for identifiability, we fix $\gamma_0 = 1$ and require $\boldsymbol{x}_i^T \boldsymbol{\gamma} > 0$ for $i = 1, \cdots, n$. Following the methods discussed in Section 2.1, we still model the error distribution by a DPM of logistic distributions mixing over both the location and scale parameters.

The model is specified as below.

$$
\begin{aligned}
& y_i|\tau_i,\sigma_i,\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{x}_i \sim \text{Logistic}(\boldsymbol{x}_i^T\boldsymbol{\beta}-\tau_i\boldsymbol{x}_i^T\boldsymbol{\gamma},\sigma_i\boldsymbol{x}_i^T\boldsymbol{\gamma}) \text{ with } \boldsymbol{x}_i^T\boldsymbol{\gamma}>0, \\
& \qquad i=1,\cdots,n, \\
& \tau_i,\sigma_i|P \sim P, \quad i=1,\cdots,n, \\
& P|\alpha,G \sim DP(\alpha,G), \\
& G(\tau,\sigma)=\text{Logistic}(\tau|-\sigma\log\lambda,\sigma)\cdot\text{Inv-Gamma}(\sigma|c,d), \\
& \beta_i \overset{i.i.d.}{\sim} \text{N}(0,\nu), \quad i=0,\cdots,m, \\
& \gamma_i \overset{i.i.d.}{\sim} \text{N}(0,\nu), \quad i=0,\cdots,m, \\
& \alpha \sim \text{Gamma}(a_1,b_1), \\
& d \sim \text{Gamma}(a_2,b_2).
\end{aligned}
\tag{4.2}
$$

As discussed in Section 2.2, to satisfy the quantile constraint, we need to make some adjustment to the posterior samples for inference of the intercept. When the heteroscedasticity is modelled in this multiplicative fashion in 4.1, we can still make simple adjustment to achieve correct inference. The difference is that we need to adjust not only the intercept, but also other regression coefficients. The detail is given in Proposition 4.1.1 in the following where the notations are the same as in Section 2.2. We still consider a simple case $Y=\beta_0+\beta_1 x+(1+\gamma x)\epsilon$ with $Q_\epsilon(p)=0$. Consider two models,

(A') $\frac{y_i-\beta_0-\beta_1 x_i}{1+\gamma x_i}|\beta_0,\beta_1,\gamma,x_i \sim F$ with pdf f_F for $i=1,\cdots,n$, $\beta_0\sim\pi_1$, $\beta_1\sim\pi_2$, $\gamma\sim\pi_4$ and $F\sim\Lambda$;

(B') $\frac{y_i-\beta_0-\beta_1 x_i}{1+\gamma x_i}|\beta_0,\beta_1,\gamma,x_i \sim F^*$ with pdf f_{F^*} for $i=1,\cdots,n$, $\beta_0\sim\pi_1$, $\beta_1\sim\pi_2$, $\gamma\sim\pi_4$ and $F^*\sim\Lambda^*$.

Let $E^{(M)}(\cdot|\mathbf{x},\mathbf{y})$ and $Var^{(M)}(\cdot|\mathbf{x},\mathbf{y})$ denote the posterior mean and variance under model (M), respectively.

Proposition 4.1.1 *If* $\pi_1(\beta_0) \propto 1$ *and* $\pi_2(\beta_1) \propto 1$,

(1) the posterior distribution of γ *in* (A') *is the same as that in* (B')*;*

(2) $E^{(B')}(\beta_0|\mathbf{x},\mathbf{y}) = E^{(A')}(\beta_0|\mathbf{x},\mathbf{y}) - E^{(A')}(q_F|\mathbf{x},\mathbf{y})$*;*

(3) $Var^{(B')}(\beta_0|\mathbf{x},\mathbf{y}) = Var^{(A')}(\beta_0 - q_F|\mathbf{x},\mathbf{y})$*;*

(4) $E^{(B')}(\beta_1|\mathbf{x},\mathbf{y}) = E^{(A')}(\beta_1|\mathbf{x},\mathbf{y}) - E^{(A')}(\gamma q_F|\mathbf{x},\mathbf{y})$*;*

(5) $Var^{(B')}(\beta_1|\mathbf{x},\mathbf{y}) = Var^{(A')}(\beta_1 - \gamma q_F|\mathbf{x},\mathbf{y})$.

These results follow from repeatedly applying change of variables and using the condition $\pi_1(\beta_0) \propto 1$. The detail of the proof is given in Appendix A.2.

Again Proposition 4.1.1 can be easily extended to the case with more covariates. Also in practice the improper prior for β_0 and β_1 can be replaced by a normal prior with a large variance. Therefore, Proposition 4.1.1 provides a guideline for the adjustments required in the MCMC inference.

As for the MCMC posterior inference, we only need to replace the residuals $y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}$ in Section 2.3 by $\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\boldsymbol{x}_i^T\boldsymbol{\gamma}}$. And we update $\boldsymbol{\gamma}$ by the Metropolis-Hastings algorithm with a reject step that rejects when $\boldsymbol{x}_i^T\boldsymbol{\gamma} \leq 0$ for some $1 \leq i \leq n$. In our experiments, this probability of this rejection is usually less than 0.1. Also after getting the samples we need to perform the adjustment suggested in Proposition 4.1.1.

4.2 Simulation study

We focus on the three DPM-based methods that extend DPMU, DPMMN and DPML with the heteroscedasticity modelled as in Section 4.1, which we call DPMUH, DPMMNH and DPMLH, respectively. We implemented all the above DPM-based methods using

the R package **Rcpp**. We still generate data according to the designs in Section 2.4 and compare DPMLH, DPMMNH and DPMUH with DPML, DPMMN and DPMU.

4.2.1 Ordinary designs

After looking into convergence diagnostics such trace plots and autocorrelation plots, we have 25,000 MCMC samples simulated and 5,000 as burn-in in each case. We set the thinning parameter to be 5. All the methods are evaluated by the MSE and PCL as in Section 2.4.

For the models DPMLH, DPMMNH and DPMUH, we set the hyperparameters the same as those in DPML, DPMMN and DPMU in Section 2.4.

For each method and each design, we report the mean and standard error of the 200 MSE's, average coverage probability of 90% interval for the slope parameters and the mean and standard error of the 200 PCL's. The results are summarized in Tables 4.1, 4.2 and 4.3.

We also get the mean square error and coverage probability for each regression coefficients including the intercept. All the detailed results can be find in the Appendix A.3.

In terms of MSE of the regression coefficients, DPMLH and DPMMNH are the best performer among all methods when the data are heteroscedastic as in Design 4. For the scenarios without heteroscedasticity, DPMLH and DPMMNH have similar performance as DPML and DPMMN for $p = 0.5$, but there is a significant decrease in their performance for $p = 0.9$. However, compared with FQR and BALD, DPMLH and DPMMNH still have significantly smaller MSEs in most cases.

In terms of coverage probability of 90% credible or confidence intervals, DPMLH and DPMMNH both have coverage probabilities close to 0.9 when there is no heteroscedas-

ticity. And DPMLH and DPMMNH have credible interval longer than DPML and DPMMN. But when there is heteroscedasticity as in Design 4, modelling heteroscedasticity explicitly dramatically improves the coverage probability of DPM-based methods.

As for the predictive check loss, due to overfitting DPMLH and DPMMNH do not perform as well as DPML and DPMMN when there is no heteroscedasticity, but DPMLH and DPMMNH are the best performers in the presence of heteroscedasticity.

Overall, when the data are heteroscedastic, DPMLH and DPMMNH have the best performance. In practice, we may first perform various statistical tests for heteroscedasticity [13, 47, 48, 86, 114], then decide whether to model heteroscedasticity. Testing the existence of heteroscedasticity deserves further research.

4.2.2 Robustness to outliers

We still consider the two scenarios with contamination as in Section 2.4.2. The results are summarized in Table 4.4, 4.5 and 4.6. The details about the intercept estimation and the simulation result for a smaller sample size $n = 100$ can be found in Appendix A.3.

DPMLH and DPMMNH have similar performance in most cases excepts for the extreme quantiles $p = 0.9$ with 10% contamination. In this case, DPMLH has coverage probabilities closer to 0.9 and shorter intervals, while DPMMNH has smaller predictive check loss. The latter is due to the fact that 10% of outliers make the estimation of the intercept totally unreliable while the intercept has a big impact on the calculation of the check loss..

Overall, if we over-parametrize the model with unnecessary parameters to account for non-existing heteroscedasticity, the gain of robustness will be compromised.

4.3 Real data study

In this section we apply our DPMLH model to analyze two real data sets – the corrected Boston housing data and the body mass index growth chart data.

4.3.1 Corrected Boston housing data

In this section, we apply our methods to analyze the corrected Boston Housing data, which was first studied in [51]. As in [56, 72, 116], the data set is a corrected version of the original data set, corrected for a few minor errors and augmented with the latitude and longitude of the observations. The data set is available in the **spdep** package in R [87]. There are 506 rows and 20 columns in the data set, with each row corresponding to one observation. The response variable is the log-transformed corrected median value of owner-occupied housing in USD 1000 ($LCMEDV$). We consider 15 predictors including point longitudes in decimal degrees (LON), point latitudes in decimal degrees (LAT), per capita crime ($CRIM$), proportions of residential land zoned for lots over 25000 square feet per town (ZN), proportions of non-retail business acres per town ($INDUS$), a factor indicating whether tract borders Charles River ($CHAS$), nitric oxides concentration (parts per 10 million) per town (NOX), average numbers of rooms per dwelling (RM), proportions of owner-occupied units built prior to 1940 (AGE), weighted distances to five Boston employment centres (DIS), index of accessibility to radial highways per town (RAD), full-value property-tax rate per USD 10,000 per town (TAX), pupil-teacher ratios per town ($PTRATIO$), transformed African American population proportion (B) and percentage values of lower status population ($LSTAT$).

As suggested in [56], we choose $\log(CRIM)$, $CHAS$, NOX^2, RM^2, AGE, $\log(DIS)$, $\log(RAD)$, TAX, $PTRATIO$, B and $\log(LSTAT)$ as the predictor variables and let

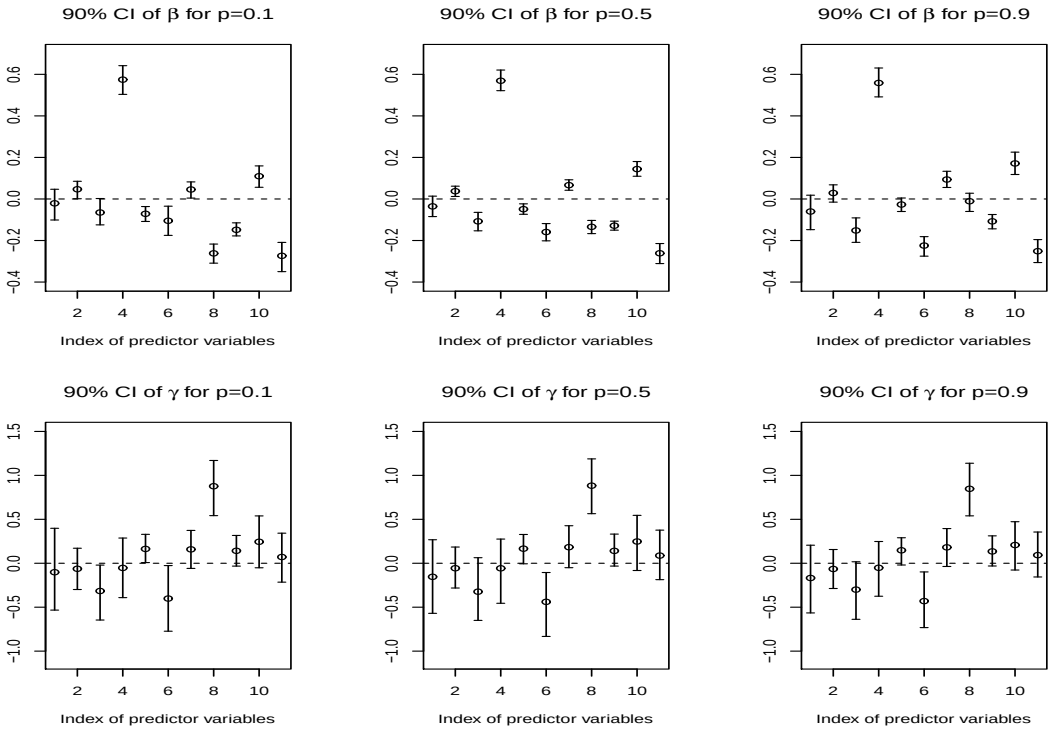
$\log(CMEDV)$ be the response variable. We also standardize all the predictor variables to make their values between -1 and 1. We are interested in three quantiles $p = 0.1, 0.5, 0.9$ and fit the data by DPMLH. Figure 4.3.1 plots the 90% credible interval for the regression coefficients and the γ's which model the heteroscedasticity. The plots (lower panel) for γ's indicate the presence of heteroscedasticity, especially for the variable TAX. As TAX increases, the variance of the housing value increases. The plots (upper panel) for the regression coefficients show that the variable RM^2 has a large positive impact on the housing value across three quantiles, while the variable $\log(LSTAT)$ has a large negative impact consistently across three quantiles. Also an interesting observation is that the negative impacts of NOX^2 and $\log(DIS)$ on the housing value get larger for more expensive houses.

4.3.2 Growth chart of body mass index (BMI)

The National Centre for Health Statistics has been conducting a national health and nutrition examination (NHANES) survey annually since 1999. The survey data are released every two years. We study the most recent available data set NHANES 2011-2012 available at `http://wwwn.cdc.gov/nchs/nhanes/search/nhanes11_12.aspx`. We are interested in how the body mass index (kg/m^2) changes with the age (in years) for the respondents with age from 2 to 20. We combine the body measurement data file with the demographic variables data file according to the respondent sequence number. After removing the respondents with missing values (29 females and 24 males), there are $n_1 = 1,700$ females and $n_2 = 1,778$ males with age from 2 to 20.

We fit a polynomial regression [81] for males and females separately. For females, let Y_i and x_i with $i = 1, \cdots, n_1$ denote the BMI and age, respectively. We first normalize

Figure 4.1. The 90% credible intervals for the regression coefficients (first row) and the γ (second row) of the correct Boston housing data for $p = 0.1, 0.5, 0.9$. Number 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 correspond to $\log(CRIM)$, $CHAS$, NOX^2, RM^2, AGE, $\log(DIS)$, $\log(RAD)$, TAX, $PTRATIO$, B and $\log(LSTAT)$.



them such that $-1 \leq x_i \leq 1$ for $i = 1, \cdots, n_1$. Then we transform x_i's into orthogonal polynomials up to order 4 as described in [52], for all $x \in \{x_1, \cdots, x_{n_1}\}$,

$$h_{r+1}(x) = 2(x - a_{r+1})h_r(x) - b_r h_{r-1}(x),$$

where $h_0(x) = 1$, $h_1(x) = 2(x - a_1)$,

$$a_{r+1} = \frac{\sum_{i=1}^{n_1} x_i h_r^2(x_i)}{\sum_{i=1}^{n_1} h_r^2(x_i)}, \quad b_r = \frac{\sum_{i=1}^{n_1} h_r^2(x_i)}{\sum_{i=1}^{n_1} h_{r-1}^2(x_i)}, \quad b_0 = 0, \quad a_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i.$$

Then we have the model

$$Y_i = \mathbf{h}(x_i)^T \boldsymbol{\beta} + \left(\mathbf{h}(x_i)^T \boldsymbol{\gamma}\right) \epsilon_i,$$

where $Q_{\epsilon_i}(p) = 0$ and $\mathbf{h}(x_i) = (h_0(x_i), h_1(x_i), h_2(x_i), h_3(x_i), h_4(x_i))^T$, $i = 1, \cdots, n_1$. The error ϵ_i is modelled by DPMLH. The same model is also fit for males.

Following [67], the quantiles of interest are taken as $p = 0.03, 0.05, 0.1, 0.25, 0.5, 0.75, 0.85, 0.95$ and 0.97. The BMI growth charts are plotted in Figure 4.3. The plots of $\mathbf{h}(x_i)^T\boldsymbol{\gamma}$ against x_i are given in Figure 4.2. Since these curves are very similar across all quantiles, we plot their average. Also the 95% credible intervals for the regression coefficients are plotted in Figures 4.5 and 4.6. It is evident that there is heteroscedasticity in the data set, because the curves in Figure 4.2 are not horizontal lines. And the heteroscedasticity has an interesting pattern. For females 2-11 and 17-20 years old, the variance of the BMI increases with age. For females 11-17 years old, the variance of the BMI decreases with age. Males have a similar but less obvious pattern. As for the BMI growth chart, for lower quantiles, there are minor difference between males and females, while for the higher quantile the BMI growth curve of females has larger curvature. Overweighted females will have a sudden increase of BMI at the age 18, while the BMI of overweighted males has a more steady increasing trend. Also by comparing Figure 4.3 with Figure 4.4, the BMI growth charts reported by Centers for Disease Control and

Prevention (CDC) in [67], the upper quantile curves increase significantly, which can be interpreted as a warning of the obesity for young people or an indicator that young people are now getting better nutrition. However, the median curve in Figure 4.4 are very similar to that in Figure 4.3, as a result, obesity is more likely to be the factor that accounts for the increase of higher quantile curves. Besides, the lower quantile curves in Figure 4.4 are even a bit higher than those in Figure 4.3, especially for females. This may be explained by the obsessive pursuit of thinness among some teenagers that may expose them to malnutrition.

Figure 4.2. Heteroscedasticity for males (left) and females (right) 2-20 years old.

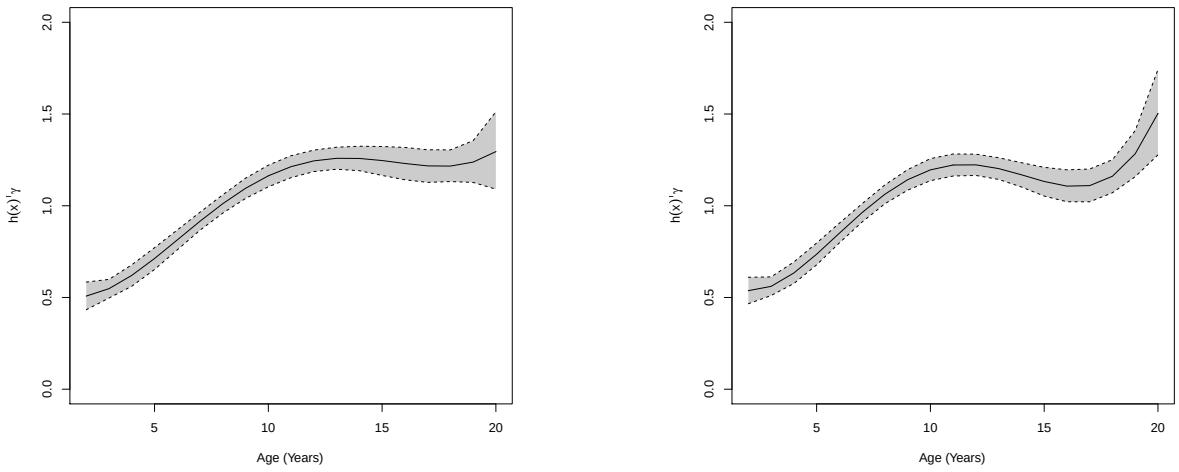


Figure 4.3. BMI growth chart for males (left) and females (right) 2-20 years old.

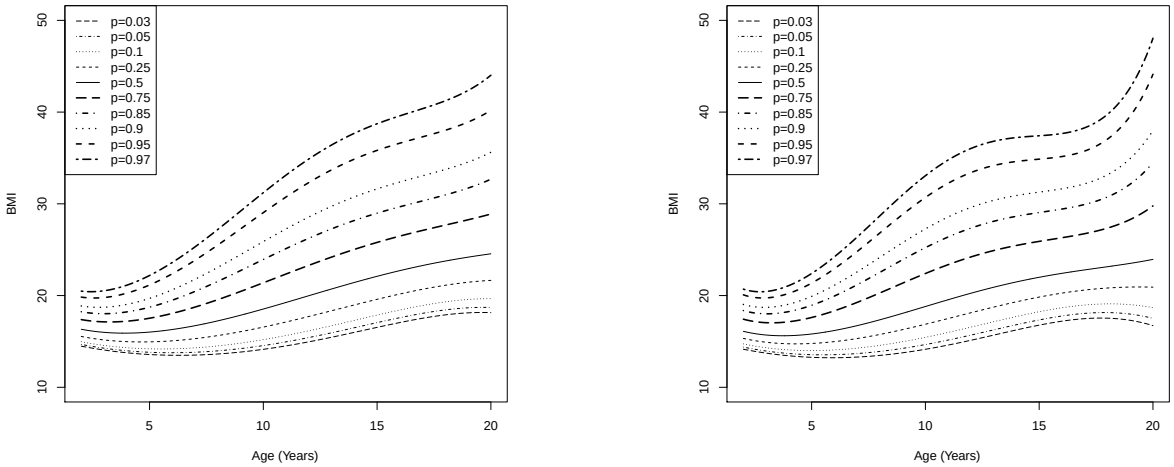


Figure 4.4. CDC BMI growth chart for males (left) and females (right) 2-20 years old.

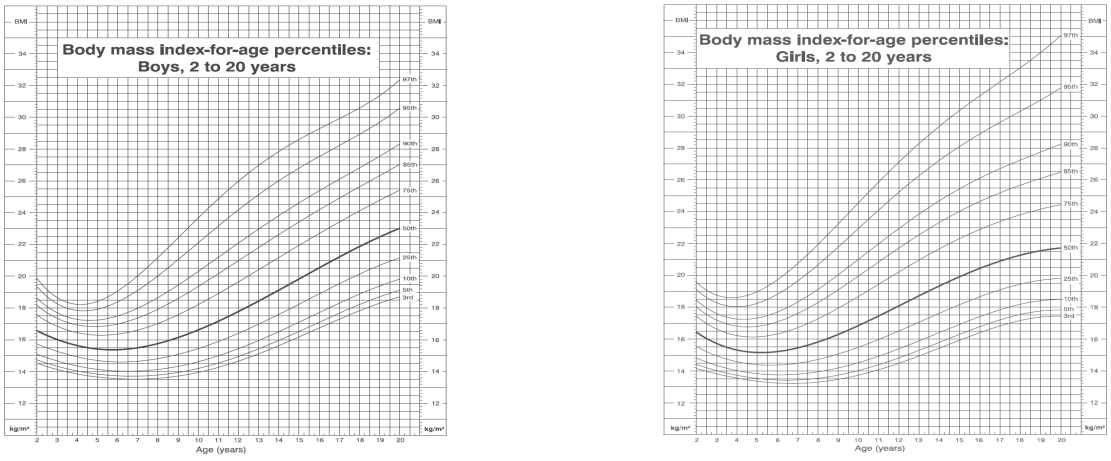


Table 4.1
Average MSE for the regression coefficients (except the intercept) for $p = 0.5$ and $p = 0.9$. MSE is reported as $100 \times$ average ($100 \times$ standard error) over the 200 simulated data sets for each design.

p	Design	DPMUH	DPMMNH	DPMLH
0.5	1	1.33	1.13	**1.12**
		(0.09)	(0.08)	(0.08)
0.5	2	2.11	2.03	**1.99**
		(0.16)	(0.14)	(0.14)
0.5	3	2.08	**1.68**	1.77
		(0.19)	(0.15)	(0.16)
0.5	4	3.01	2.38	**2.33**
		(0.28)	(0.23)	(0.22)
0.5	5	1.17	**0.94**	0.95
		(0.07)	(0.06)	(0.06)
0.5	6	2.57	1.99	**1.97**
		(0.20)	(0.13)	(0.13)
0.9	1	2.42	2.30	**2.07**
		(0.20)	(0.18)	(0.16)
0.9	2	**7.53**	10.08	8.69
		(0.60)	(0.76)	(0.63)
0.9	3	4.77	5.07	**4.61**
		(0.38)	(0.39)	(0.37)
0.9	4	5.89	4.35	**3.93**
		(0.66)	(0.45)	(0.41)
0.9	5	1.82	1.69	**1.68**
		(0.13)	(0.12)	(0.11)
0.9	6	5.87	5.75	**4.99**
		(0.44)	(0.44)	(0.39)

Table 4.2
Average coverage probability (CP) of 90% credible or confidence intervals of the regression coefficients (except the intercept) for $p = 0.5$ and $p = 0.9$. The average length of intervals is also reported.

p	Design		DPMUH	DPMMNH	DPMLH
0.5	1	CP	0.85	**0.90**	0.89
		Length	0.33	0.34	0.34
0.5	2	CP	0.89	**0.90**	0.89
		Length	0.46	0.48	0.46
0.5	3	CP	0.83	**0.87**	**0.87**
		Length	0.38	0.39	0.39
0.5	4	CP	0.87	**0.89**	0.87
		Length	0.55	0.50	0.48
0.5	5	CP	0.85	**0.89**	**0.89**
		Length	0.30	0.30	0.30
0.5	6	CP	0.82	**0.86**	**0.86**
		Length	0.43	0.42	0.42
0.9	1	CP	0.79	**0.89**	**0.89**
		Length	0.41	0.51	0.48
0.9	2	CP	0.75	**0.91**	0.86
		Length	0.65	1.10	0.94
0.9	3	CP	0.71	**0.89**	0.85
		Length	0.49	0.72	0.63
0.9	4	CP	0.77	**0.94**	**0.94**
		Length	0.62	0.76	0.69
0.9	5	CP	0.86	**0.91**	**0.91**
		Length	0.39	0.45	0.43
0.9	6	CP	0.71	**0.88**	0.86
		Length	0.54	0.74	0.66

Table 4.3
Predictive check loss for $p = 0.5$ and $p = 0.9$. Standard error is reported in the parenthesis.

p	Design	DPMUH	DPMMNH	DPMLH
0.5	1	4135.55	**4121**	4122.35
		(4.74)	(3.63)	(3.72)
0.5	2	6233.65	6231.86	**6225.88**
		(6.61)	(6.51)	(6.16)
0.5	3	5116.2	**5102.05**	5105.86
		(7.65)	(6.79)	(7.02)
0.5	4	4355.48	4345.1	**4344.78**
		(4.04)	(3.3)	(3.26)
0.5	5	4141.59	**4110.58**	4113.29
		(9.14)	(6.15)	(6.27)
0.5	6	5600.77	5577.63	**5576.98**
		(7.36)	(5.49)	(5.46)
0.9	1	1859.68	1825.08	**1821.26**
		(5.96)	(3.65)	(3.74)
0.9	2	4274.5	**3920.25**	3938.43
		(20.49)	(7.47)	(9.08)
0.9	3	2784.22	2670.11	**2656.35**
		(11.35)	(5.09)	(4.58)
0.9	4	1966.49	1952.69	**1949.01**
		(4.01)	(2.62)	(2.46)
0.9	5	1901.11	1862.44	**1854.71**
		(7.74)	(6.54)	(5.53)
0.9	6	3064.14	3002.98	**2983.15**
		(9.65)	(6.01)	(4.94)

Table 4.4
Mean square error for the regression coefficients when there are outliers. The standard error is reported in the parenthesis.

Contamination proportion	p	Coefficient	DPMMNH	DPMLH
5%	0.5	β_1	**0.176**	0.198
			(0.014)	(0.016)
5%	0.5	β_2	**0.199**	0.208
			(0.02)	(0.022)
5%	0.9	β_1	**0.11**	0.116
			(0.012)	(0.013)
5%	0.9	β_2	**0.122**	0.132
			(0.013)	(0.014)
10%	0.5	β_1	0.157	**0.156**
			(0.014)	(0.014)
10%	0.5	β_2	0.208	**0.204**
			(0.023)	(0.023)
10%	0.9	β_1	0.026	**0.021**
			(0.003)	(0.002)
10%	0.9	β_2	0.023	**0.019**
			(0.003)	(0.002)

Table 4.5
Lengths of 90% credible or confidence intervals and coverage probabilities (CP) for regression coefficients when there are outliers.

Contamination proportion	p	Coefficient		DPMMNH	DPMLH
5%	0.5	β_1	CP	0.93	**0.895**
			Length	0.146	0.146
5%	0.5	β_2	CP	0.92	**0.9**
			Length	0.146	0.145
5%	0.9	β_1	CP	**0.91**	0.88
			Length	0.119	0.112
5%	0.9	β_2	CP	0.915	**0.895**
			Length	0.119	0.112
10%	0.5	β_1	CP	**0.935**	0.96
			Length	0.146	0.147
10%	0.5	β_2	CP	**0.895**	**0.905**
			Length	0.146	0.146
10%	0.9	β_1	CP	1	**0.88**
			Length	0.094	0.048
10%	0.9	β_2	CP	0.995	**0.925**
			Length	0.094	0.048

Table 4.6
Predictive check loss when there are outliers. The standard error is reported in the parenthesis

Contamination proportion	p	DPMMNH	DPMLH
5%	0.5	3979.84	**3979.69**
		(1.201)	(1.175)
5%	0.9	2113.19	**1887.68**
		(56.64)	(2.537)
10%	0.5	4018.62	**4014.58**
		(2.302)	(2.108)
10%	0.9	**5338.42**	5748.4
		(39.00)	(4.044)

Figure 4.5. 95% credible intervals for the regression coefficients for males 2-20 years old for various quantiles.

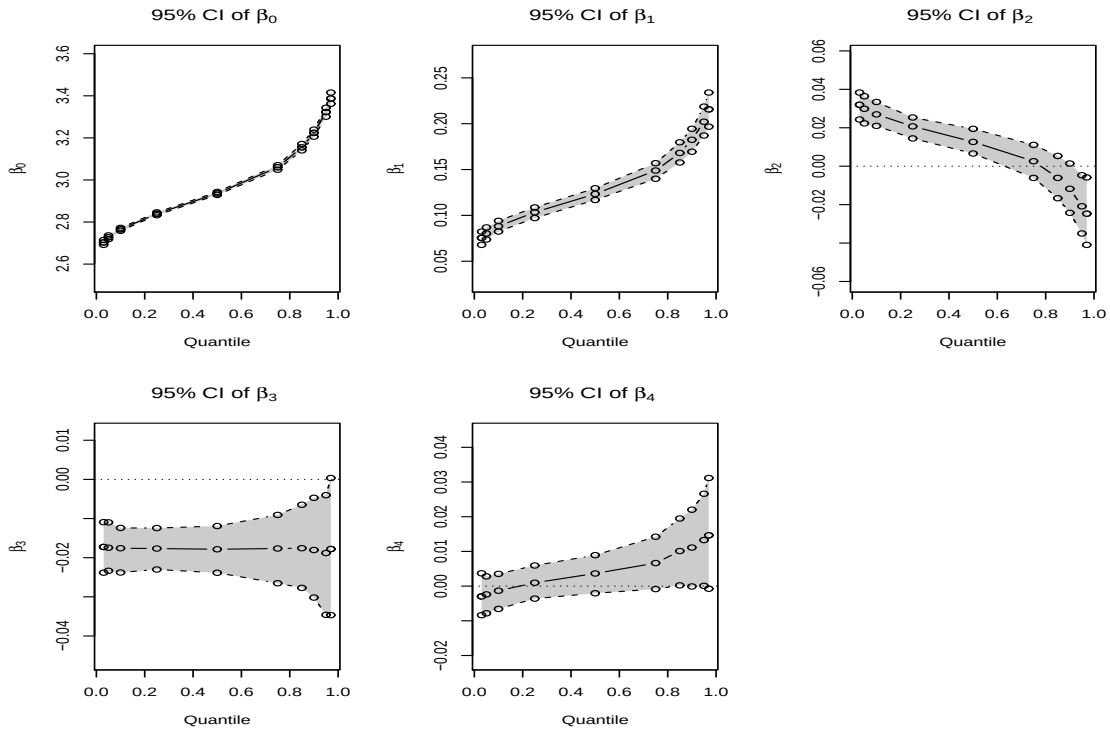
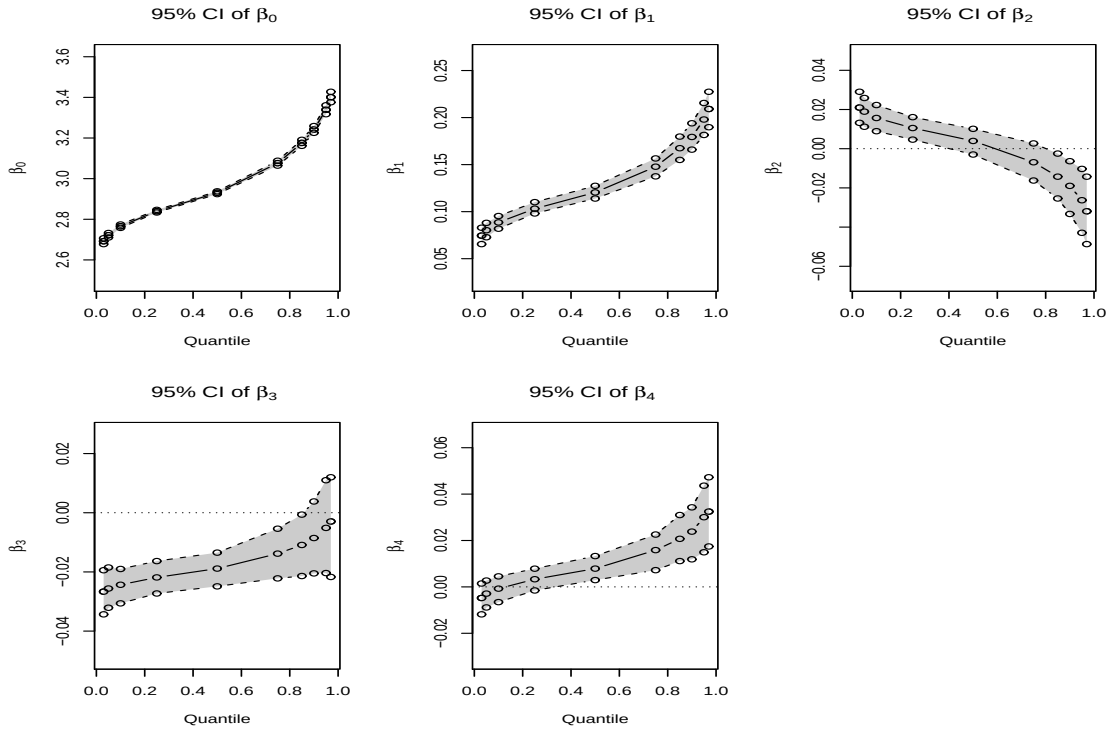


Figure 4.6. 95% credible intervals for the regression coefficients for females 2-20 years old for various quantiles.



5. Quantile regression for longitudinal data

A longitudinal study usually tracks the same variables on the same subjects over a time period. Longitudinal data occur in a wide range of studies like clinical trials and panel studies [24, 105]. For example, to investigate the associations between exposure to suspected causes of disease and subsequent morbidity or mortality, we need to follow up with the same group of participants over a long period and record their exposure to the causes as well as their health status at multiple follow-up times. The repeated measurements in a longitudinal study are correlated within subjects, thus special statistical techniques are required for valid analysis.

In this chapter, we extend the ideas in Chapters 2 and 4 to handle longitudinal data. Sections 2.2 and 4.1 introduced how to adjust the posterior sample to satisfy the quantile constraint. This implies that we can completely ignore the quantile constraint when we choose the kernel densities. This opens the door to a variety of DPM-based regression models. Moreover, as discussed in Remark 3.6, the posterior consistency holds for the DPM of normal distributions (DPMN) as in (3.25). In this chapter, we analyze longitudinal data using Bayesian quantile regression using the DPMN model.

5.1 Quantile regression for longitudinal data

There are considerable efforts in the frequentist literature to extend quantile regression to handle longitudinal data. [60] employed regularization methods to estimate quantile regression models for longitudinal data. [58] established sufficient conditions for

consistency and asymptotic normality for the quantile regression estimator in the presence of individual effects. [35] proposed a method to handle location-shift random effects by modelling the error distribution by the ALD.

On the other hand, some Bayesian quantile regression methods are also developed to analyze longitudinal data. [3, 120] developed parametric Bayesian approaches using the ALD to model the error distribution. [89] proposed a DPM-based method, where the kernel density is specified as a two-component normal mixture.

We propose a novel nonparametric Bayesian method for quantile regression via the DPM of normal distributions. Although the normal density is a popular choice for the kernel in mixture models, no existing literature used the normal density as the kernel of DPM for quantile regression. This is largely because the normal distribution has no closed-form quantile function, hence it is hard to guarantee that the mixture models with normal kernels satisfy the quantile constraint. [89] overcomes this difficulty by making the kernel as a mixture of two normal distributions. However, having the kernel itself as a mixture increases the complexity of the model and reduces the computational efficiency. Instead of requiring the kernel to meet the quantile constraint, we adjust the DPM of normal distributions by a location shift to guarantee the quantile constraint. Therefore, our model is both simple and flexible. And by properly choosing the base measure of the Dirichlet process, we have full conjugacy and a computationally efficient algorithm for posterior inference.

5.2 The model

Consider the repeated measurement data in the form $(\boldsymbol{x}_{it}, y_{it})$ for $i = 1, \cdots, n$ and $t = 1, \cdots, T$, where $\boldsymbol{x}_{it}$ is a column m-vector and y_{it} is the t-th measurement of a

continuous univariate response on the i-th subject. In the sequel, the quantile of interest is always the p-th quantile. Let $\phi(x)$ denote the pdf of the standard normal distribution.

We consider the model

$$Y_{it} = \theta_i + \boldsymbol{x}_{it}^T\boldsymbol{\beta} + (1 + \boldsymbol{x}_{it}^T\boldsymbol{\gamma})\epsilon_{it}, \quad \text{for } i = 1, \cdots, n \text{ and } t = 1, \cdots, T, \tag{5.1}$$

where θ_i is the individual effect for the i-th subject and $Q_\epsilon(p) = 0$. The heteroscedasticity is modelled by multiplying a linear term to the error as in [53, 89]. As in Chapter 4, for identifiability, we require $1 + \boldsymbol{x}_{it}^T\boldsymbol{\gamma} > 0$ for $i = 1, \cdots, n$ and $t = 1, \cdots, T$.

We summarize our proposed model as follows.

$$\begin{aligned}
&y_{it}|\mu_{it}, \sigma_{it}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{x}_{it}, \theta_i \sim \mathrm{N}\left(\theta_i + \boldsymbol{x}_{it}^T\boldsymbol{\beta} + \mu_{it}\boldsymbol{x}_{it}^T\boldsymbol{\gamma}, \sigma_{it}^2\left(\boldsymbol{x}_{it}^T\boldsymbol{\gamma}\right)^2\right) \\
&\qquad \text{with } \boldsymbol{x}_{it}^T\boldsymbol{\gamma} > 0, \quad i = 1, \cdots, n, \quad t = 1, \cdots, T, \\
&\mu_{it}, \sigma_{it}^2|P \sim P, \quad i = 1, \cdots, n, \quad t = 1, \cdots, T, \\
&P|\alpha, G \sim DP(\alpha, G), \\
&G(\mu, \sigma^2) = \mathrm{N}(\mu|0, \sigma^2) \cdot \text{Inv-Gamma}(\sigma^2|c, d), \\
&\theta_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \nu), \quad i = 0, \cdots, n, \\
&\beta_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \nu), \quad i = 0, \cdots, m, \\
&\gamma_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \nu), \quad i = 0, \cdots, m, \\
&\alpha \sim \text{Gamma}(a_1, b_1), \\
&d \sim \text{Gamma}(a_2, b_2),
\end{aligned} \tag{5.2}$$

with hyper-parameters c, ν, a_1, a_2, b_1 and b_2.

In model (5.2), the distribution of ϵ_{it} is modelled by the DPM of the normal kernels

$$f_{\epsilon_{it}}(z) = \int \frac{1}{\sigma}\phi\left(\frac{x - \mu}{\sigma}\right) dP(\mu, \sigma^2), \quad P \sim DP(\alpha, G). \tag{5.3}$$

Note that the base measure G is the normal-inverse-gamma distribution and hence conjugate to the our normal kernel, which greatly simplifies the posterior inference. As in

Chapter 4, the error distribution violates the quantile constraint. In the next section, we propose a simple adjustment to get valid inference for the regression coefficients as well as the random effects.

5.3 Adjustment

In this subsection we will study how the quantile constraint affects the inference. And a simple adjustment can be made to correct the estimation for the regression coefficients as well as the random effects. Similar as in Sections 2.2 and 4.1, we first introduce some notations to allow discussing the problem in a more general setting. Still consider a quantile regression model with a univariate covariate, $Y_{it} = \theta_i + \beta x_{it} + (1 + x_{it}\gamma)\epsilon_{it}$ for $i = 1, \cdots, n$ and $t = 1, \cdots, T$, with the quantile constraint $Q_{\epsilon_{it}}(p) = 0$. We assume prior independence between θ_i's, β, γ and let $\pi_1(\theta_i)$, $\pi_2(\beta)$ and $\pi_3(\gamma)$ denote the independent priors for θ_i's, β and γ, respectively. We also assume π_1 and π_2 are both supported in $(-\infty, \infty)$. Let Λ denote any probability measure over the space of probability measures which are absolutely continuous with respect to the Lebesgue measure. Consider the model

$$(A'') \quad \frac{y_{it} - \theta_i - \beta x_{it}}{1 + \gamma x_{it}} | \theta_i, \beta, \gamma, x_{it} \sim F \text{ with pdf } f_F \text{ for } i = 1, \cdots, n,\ t = 1, \cdots, T,$$

$$\theta_i \sim \pi_1, \beta \sim \pi_2, \gamma \sim \pi_3 \text{ and } F \sim \Lambda.$$

As there is no constraint on Λ, the quantile constraint may be violated in model (A''). For each $f_F(z)$, define q_F such that $f_F(z - q_F)$ satisfies the quantile constraint. The fact that the quantile constraint is on the location parameter of f_F guarantees the existence of such q_F is directly from . Then we can define a random probability measure Λ^* based

on Λ. We say $F^* \sim \Lambda^*$ if and only if there exists $F \sim \Lambda$ such that the pdf of F^* is given by $f_{F^*}(z) = f_F(z - q_F)$. So we have another model

$$(B'') \quad \frac{y_{it} - \theta_i - \beta x_{it}}{1 + \gamma x_{it}} | \theta_i, \beta, \gamma, x_{it} \sim F^* \text{ with pdf } f_{F^*} \text{ for } i = 1, \cdots, n,\ t = 1, \cdots, T,$$

$$\theta_i \sim \pi_1, \beta \sim \pi_2, \gamma \sim \pi_3 \text{ and } F^* \sim \Lambda^*.$$

By the definition of F^*, the quantile constraint is satisfied in model (B''). Next we will show that there is a simple relation between the posterior inference in model (A'') and (B'').

Let $E^{(M)}(\cdot|\mathbf{x}, \mathbf{y})$ and $Var^{(M)}(\cdot|\mathbf{x}, \mathbf{y})$ denote the posterior mean and variance under model (M), respectively.

Proposition 5.3.1 *If $\pi_1(\theta_i) \propto 1$ and $\pi_2(\beta) \propto 1$, then*

(1) the posterior distribution of γ in (A'') is the same as that in (B'');

(2) $E^{(B'')}(\theta_i|\mathbf{x}, \mathbf{y}) = E^{(A'')}(\theta_i|\mathbf{x}, \mathbf{y}) - E^{(A'')}(q_F|\mathbf{x}, \mathbf{y})$;

(3) $Var^{(B'')}(\theta_i|\mathbf{x}, \mathbf{y}) = Var^{(A'')}(\theta_i - q_F|\mathbf{x}, \mathbf{y})$;

(4) $E^{(B'')}(\beta|\mathbf{x}, \mathbf{y}) = E^{(A'')}(\beta|\mathbf{x}, \mathbf{y}) - E^{(A'')}(\gamma q_F|\mathbf{x}, \mathbf{y})$;

(5) $Var^{(B'')}(\beta|\mathbf{x}, \mathbf{y}) = Var^{(A'')}(\beta - \gamma q_F|\mathbf{x}, \mathbf{y})$.

Proof The results follow from repeatedly applying change of variables and using the conditions$\pi_1(\theta_i) \propto 1$ and $\pi_2(\beta) \propto 1$. The proof is essentially the same as that for Proposition 4.1.1. ■

Proposition 5.3.1 suggests that with MCMC sample $\{\theta_1^{(t)}, \cdots, \theta_n^{(t)}, \beta^{(t)}, q_F^{(t)}, (\gamma q_F)^{(t)}\}_{t=1}^T$ from model (A''), to get valid inference for $\{\theta_i\}_{i=1}^n, \beta$ we should work with the adjusted sample $\{\theta_1^{(t)} - q_F^{(t)}, \cdots, \theta_n^{(t)} - q_F^{(t)}, \beta^{(t)} - (\gamma q_F)^{(t)}\}_{t=1}^T$.

Proposition 5.3.1 is ready to be extended to the case with more covariates. And in practice, we may replace the improper prior for β by a normal distribution with a large variance ν. The estimation error for regression coefficients using the posterior mean is bounded by $\frac{C}{\nu}$, where C is a constant independent of ν. So practically our model (5.3) can be treated as one example of model (A''). After making the adjustment suggested in Proposition 5.3.1, we get the same posterior inference as the model which employs a location shift of the mixture to satisfy the quantile constraint in the same fashion as in model (B'').

5.4 Posterior inference

In this section, we provide a MCMC algorithm for the posterior inference of model (5.2). As discussed in the Section 5.3, we need to use the adjusted MCMC sample for inference. We provide a Gibbs sampler here. Each iteration of the Markov chain updates (i) the precision parameter α, (ii) the scale parameter d in the base measure, (iii) the regression coefficients $\boldsymbol{\beta}$, (iv) the $\boldsymbol{\gamma}$, (v) the θ's and (vi) the pairs of location and scale parameters for each sample $\{(\mu_l, \sigma_l^2)\}_{l=1}^N$, where $N = nT$. Let N^* denote the number of clusters, i.e. the number of distinct pairs in $\{(\mu_l, \sigma_l)\}_{l=1}^N$. And let $\{(\mu_j^*, \sigma_j^{2*})\}_{j=1}^{N^*}$ denote the distinct pairs.

(i) The full conditional distribution for α is hard to get directly, but as a standard trick [29], one can introduce a fictitious parameter η with prior $U(0,1)$ and update α together with η. Let $\pi_{\eta,N^*} = \frac{a_1+N^*-1}{a_1+N^*-1+N(b_1-\log(\eta))}$, then

$$\alpha|\eta, N^* \sim \begin{cases} \text{Gamma}(a_1+N^*, b_1-\log(\eta)), & \text{with probability } \pi_{\eta,N^*}; \\ \text{Gamma}(a_1+N^*-1, b_1-\log(\eta)), & \text{with probability } 1-\pi_{\eta,N^*}. \end{cases} \tag{5.4}$$

$$\eta|\alpha, N^* \sim Beta(\alpha+1, N).$$

(ii) The full conditional distribution for d is given by

$$d|N^*, \sigma_1^{2*}, \cdots, \sigma_{N^*}^{2*} \sim \text{Gamma}\left(a_2+2N^*, b_2+\sum_{j=1}^{N^*}\frac{1}{\sigma_j^{2*}}\right). \tag{5.5}$$

(iii) Let $\mathbf{X} := (\boldsymbol{x}_{11}, \cdots, \boldsymbol{x}_{1T}, \cdots, \boldsymbol{x}_{n1}, \cdots, \boldsymbol{x}_{nT})^T$, $\mathbf{V} := \text{Diag}\{\mathbf{X}\boldsymbol{\gamma}\}$,

$\mathbf{Y} := (y_{11}, \cdots, y_{1T}, \cdots, y_{n1}, \cdots, y_{nT})^T$, $\boldsymbol{\mu} := (\mu_{11}, \cdots, \mu_{1T}, \cdots, \mu_{n1}, \cdots, \mu_{nT})^T$,

$\boldsymbol{\Sigma} := \text{Diag}\{\sigma_{11}^2(\boldsymbol{x}_{11}\boldsymbol{\gamma})^2, \cdots, \sigma_{1T}^2(\boldsymbol{x}_{1T}\boldsymbol{\gamma})^2, \cdots, \sigma_{n1}^2(\boldsymbol{x}_{n1}\boldsymbol{\gamma})^2, \cdots, \sigma_{nT}^2(\boldsymbol{x}_{nT}\boldsymbol{\gamma})^2\}$,

$\boldsymbol{\Theta} := ((\theta_1\mathbf{1}_T)^T, \cdots, (\theta_n\mathbf{1}_T)^T)^T$, $\mathbf{W} := \text{Diag}\{\nu\mathbf{1}_N\}$, where $\mathbf{1}_k := (1, \cdots, 1)^T \in \mathbb{R}^k$ and Diag($\mathbf{v}$) denote the diagonal matrix with diagonal elements being $\mathbf{v}$ and off-diagonal elements being 0. Also let $\boldsymbol{\Omega} := (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X} + \mathbf{W}^{-1})^{-1}$. The full conditional distribution of $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{W}, \mathbf{V}, \boldsymbol{\Theta}, \mathbf{Y}, \mathbf{X} \sim \text{N}_m\left(\boldsymbol{\Omega}\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\boldsymbol{\Theta}-\mathbf{V}\boldsymbol{\mu}), \boldsymbol{\Omega}\right). \tag{5.6}$$

(iv) $\boldsymbol{\gamma}$ is updated by the Metropolis-Hastings algorithm. And if a $\boldsymbol{\gamma}$ makes $\boldsymbol{x}_{it}^T\boldsymbol{\gamma} <= 0$ for some $i \in \{1, \cdots, n\}$, $t \in \{1, \cdots, T\}$, we reject and regenerate $\boldsymbol{\gamma}$.

(v) Let $u_{it} := y_{it} - \boldsymbol{x}_{it}^T\boldsymbol{\beta} - \boldsymbol{x}_{it}^T\boldsymbol{\gamma}\mu_{it}$ and $v_{it} := \sigma_{it}^2\left(\boldsymbol{x}_{it}^T\boldsymbol{\gamma}\right)^2$. Also let $\mathbf{u}_i := (u_{i1}, \cdots, u_{iT})^T$ and $\mathbf{v}_i := (v_{i1}, \cdots, v_{iT})^T$. Then the full conditional distribution of θ_i is given by

$$\theta_i|\mathbf{u_i}, \mathbf{v}_i \sim \text{N}\left(\left(\sum_{t=1}^{T}\frac{1}{v_{it}}+\frac{1}{\nu}\right)^{-1}\sum_{t=1}^{T}\frac{u_{it}}{v_{it}}, \left(\sum_{t=1}^{T}\frac{1}{v_{it}}+\frac{1}{\nu}\right)^{-1}\right). \tag{5.7}$$

(vi) We split this step into two parts. First we update the unique pairs (μ_j^*, σ_j^{2*}), and then we update the cluster configuration. Let $e_{it} := \frac{y_{it}-\mathbf{x}_{it}^T\boldsymbol{\beta}-\theta_i}{\mathbf{x}_{it}^T\boldsymbol{\gamma}}$, $\mathbf{e} := \{e_{11}, \cdots, e_{1T}, \cdots, e_{n1}, \cdots, e_{nT}\}$ and $A_j := \{(i,t) : \ \mu_{it} = \mu_j^*, \sigma_{it}^2 = \sigma_j^{2*}\}$. The full conditional distribution of μ_j^* is given by

$$\mu_j^* | \mathbf{e}, \sigma_j^{2*} \sim \mathrm{N}\left(\frac{\sum\limits_{(i,t)\in A_j} e_{it}}{|A_j|+1}, \frac{\sigma_j^{2*}}{|A_j|+1} \right). \tag{5.8}$$

And the full conditional distribution for σ_j^{2*} is given by

$$\sigma_j^{2*} | \mathbf{e}, \mu_j^* \sim \text{Inv-Gamma}\left(c + \frac{1}{2} + \frac{|A_j|}{2}, d + \frac{\sum\limits_{(i,t)\in A_j} (e_{it}-\mu_j^*)^2}{2} + \frac{(\mu_j^*)^2}{2} \right). \tag{5.9}$$

Next, consider updates of the configuration of clusters, that is, which samples fall into which cluster and the number of clusters. This is the key step of MCMC sampling for DPM models. Note that in our case the base measure is a conjugate prior, thus the update of configuration is standard. Following the tradition [29, 82], let c_l, $l = 1, \cdots, N$, denote the cluster indicator of the l-th observation. Without loss of generality, assume $c_l = k$ if $(\mu_l, \sigma_l^2) = (\mu_k^*, \sigma_k^{2*})$. Let $n_{-l,c} := |\{j : 1 \le j \le N, \ j \ne l, \ c_j = c\}|$, $\boldsymbol{\mu}^* = (\mu_1^*, \cdots, \mu_{N^*}^*)^T$, $\boldsymbol{\sigma}^{2*} = (\sigma_1^{2*}, \cdots, \sigma_{N^*}^{2*})^T$ and let $\phi(z|\mu, \sigma^2)$ denote the pdf of the normal distribution with mean μ and variance σ^2. Then for $l = 1, \cdots, N$, let $(\boldsymbol{x}_{it}, y_{it})$ be the corresponding observation, we iterate the following two steps,

(1) Draw c_l according to

$$Pr(c_l = c | \mathbf{e}, \boldsymbol{\mu}^*, \boldsymbol{\sigma}^{2*}, n_{-l,c}) = \begin{cases} K n_{-l,c} \phi(e_{it} | \mu_c^*, \sigma_c^{2*}), & \text{for } 1 \le c \le N^*; \\ K \frac{\alpha d^c \Gamma(c+1/2)}{2\sqrt{\pi}\Gamma(c)\left(d + e_{it}^2/4\right)^{c+1/2}}, & \text{for } c \notin \{1, \cdots, N^*\}, \end{cases} \tag{5.10}$$

where $\Gamma(t)$ denotes the Gamma function and K is a normalizing constant.

(2) If $1 \leq c_l \leq N^*$, (μ_l, σ_l^2) is updated by $(\mu_{c_l}^*, \sigma_{c_l}^{2*})$. And if $c_l \notin \{1, \cdots, N^*\}$, (μ_l, σ_l^2) is updated by

$$\mu_l | e_{it}, \sigma_l^2 \sim \mathrm{N}\left(\frac{e_{it}}{2}, \frac{\sigma_l^2}{2}\right),$$
$$\text{and} \quad \sigma_l^2 | e_{it}, \mu_l \sim \text{Inv-Gamma}\left(c+1, d + \frac{(e_{it}-\mu_l)^2}{2} + \frac{(\mu_l)^2}{2}\right).$$

After each MCMC iteration, we need to compute q_F to perform the adjustment described in Proposition 5.3.1. Assume we have n^* clusters with cluster sizes s_j for $j = 1, \cdots, n^*$ and unique pairs $(\mu_j^*, \sigma_j^{2*})_{j=1}^{n^*}$. q_F is the p-th quantile of the mixture distribution with pdf $f(x) = \sum_{j=1}^{n^*} \frac{s_j}{n^*} \frac{1}{\sigma_j^*} \phi\left(\frac{x-\mu_j^*}{\sigma_j^*}\right)$ and is given by the equation

$$\int_{-\infty}^{q_F} \sum_{j=1}^{n^*} \frac{s_j}{n^*} \frac{1}{\sigma_j^*} \phi\left(\frac{x+\tau_j^*}{\sigma_j^*}\right) dx = p.$$

. Since given (μ_j^*, σ_j^{2*}) and any $q \in \mathbb{R}$, we are able to compute $\Phi\left(\frac{q-\mu_j^*}{\sigma_j^{2*}}\right)$ using standard libraries such as **Rcpp**, q_F can be quickly computed by a binary search.

5.5 Simulation study

Our simulation study mainly follows [58]. We have six designs:

- **Design I**: $y_{it} = \theta_i + x_{it}\beta + \epsilon_{it}^1$;
- **Design II**: $y_{it} = \theta_i + x_{it}\beta + \epsilon_{it}^2$;
- **Design III**: $y_{it} = \theta_i + x_{it}\beta + \epsilon_{it}^3$;
- **Design IV**: $y_{it} = \theta_i + x_{it}\beta + (1 + x_{it}\gamma)\epsilon_{it}^1$;
- **Design V**: $y_{it} = \theta_i + x_{it}\beta + (1 + x_{it}\gamma)\epsilon_{it}^2$;
- **Design VI**: $y_{it} = \theta_i + x_{it}\beta + (1 + x_{it}\gamma)\epsilon_{it}^3$,

where $x_{it} = 0.3\theta_i + z_{it}$, $z_{it} \overset{i.i.d.}{\sim} \chi^2_3$, $\theta_i \overset{i.i.d}{\sim} U(0,1)$, $\epsilon^1_{it} \overset{i.i.d.}{\sim} \mathrm{N}(0,1)$, $\epsilon^2_{it} \overset{i.i.d.}{\sim} \chi^2_3$ and $\epsilon^3_{it} \overset{i.i.d.}{\sim}$ Cauchy$(0,1)$. We set $\beta = \gamma = 1$. The quantiles of interest are $p = 0.1, 0.25, 0.5, 0.75$ and 0.9. The sample size has two settings, $n = 25, t = 5$ and $n = 50, t = 20$. The number of Monte Carlo repetition is 200 in all scenarios.

For model (5.2), we still set the hyper-parameters $a_1 = b_1 = 1$, $c = 2$, $\nu = 10^8$ and for a_2 and b_2, we use the empirical Bayes method. Since all σ_i's have prior mean equal to d. We want to make the prior mean of d be large enough to capture the dispersion in the data. To achieve this we set $a_2 = 1$ and $b_2 = \max_{i=1,\cdots,n} y_i - \min_{i=1,\cdots,n} y_i$.

The performance of our model is evaluated by the bias and standard deviation of the posterior mean $\hat{\beta}$ of β and the coverage probability of the 95% credible interval. The results are summarized in Tables 5.1, 5.2 and 5.3.

For the designs with no heteroscedasticity (Design I-III), the bias is small across all cases. However, for designs with heteroscedasticity (Design IV-VI), the bias is significantly amplified for the extreme quantiles, especially when the sample size is small. Similarly, the standard deviation is small for cases with no heteroscedasticity and is much larger for the extreme quantiles of heteroscedastic cases. As for the coverage probability, poor coverage happens for the extreme quantiles of the heteroscedastic cases with few observations. For example, for $p = 0.9$, $n = 25$, $T = 5$, the coverage probability is only 0.74 in Design VI.

As seen in the simulation study, for the heavy-tailed error density, the normal kernel does not perform very well, because more component are required to fit heavy-tailed densities with the normal mixtures. A potential improvement is to use t distributions as the kernel, and put a prior on the degrees of freedom. However, we no longer have conjugacy when working with the t distribution. Metropolis-Hastings algorithms and

more complicated sampling methods for the DPM are required. We will explore in this direction in future works.

Table 5.1
Average bias of $\hat{\beta}$.

p	n	T	Design I	Design II	Design III	Design IV	Design V	Design VI
0.1	25	5	-0.0069	0.016	0.0212	0.1065	0.0191	0.9993
			(0.0044)	(0.0059)	(0.0135)	(0.0194)	(0.0251)	(0.0706)
0.1	50	20	-7e-04	-2e-04	0.0164	0.0085	0.0509	0.3154
			(0.0012)	(0.0012)	(0.0044)	(0.0071)	(0.0066)	(0.0289)
0.25	25	5	-0.0049	0.0144	0.0095	0.0625	0.1966	0.1162
			(0.0035)	(0.0053)	(0.0077)	(0.0155)	(0.0267)	(0.0381)
0.25	50	20	-6e-04	-0.0017	0.0064	0.0038	-0.0217	0.0594
			(0.001)	(0.0012)	(0.0021)	(0.0058)	(0.0077)	(0.0109)
0.5	25	5	-0.0027	0.0126	0.0021	0.0061	-0.003	-0.0107
			(0.0031)	(0.006)	(0.0059)	(0.014)	(0.0356)	(0.0293)
0.5	50	20	-6e-04	-0.0045	0.0014	-0.0034	-0.088	-0.002
			(9e-04)	(0.0018)	(0.0015)	(0.0053)	(0.0107)	(0.0077)
0.75	25	5	-6e-04	0.0099	-0.0051	-0.0493	-0.4978	-0.138
			(0.0035)	(0.0085)	(0.0081)	(0.0162)	(0.0495)	(0.0401)
0.75	50	20	-5e-04	-0.0089	-0.0037	-0.0111	-0.174	-0.0646
			(0.001)	(0.0032)	(0.002)	(0.0059)	(0.0172)	(0.0108)
0.9	25	5	0.0014	0.0059	-0.0172	-0.0906	-0.9358	-1.0363
			(0.0045)	(0.0131)	(0.0141)	(0.0203)	(0.0714)	(0.071)
0.9	50	20	-5e-04	-0.0144	-0.0142	-0.0173	-0.2885	-0.334
			(0.0012)	(0.0051)	(0.0044)	(0.0072)	(0.0267)	(0.0265)

Table 5.2
Average estimated standard error of $\hat{\beta}$.

p	n	T	Design I	Design II	Design III	Design IV	Design V	Design VI
0.1	25	5	0.0636	0.1023	0.1881	0.3029	0.4987	0.9607
			(8e-04)	(0.0016)	(0.0041)	(0.0024)	(0.0057)	(0.0215)
0.1	50	20	0.0182	0.0177	0.0577	0.0954	0.0905	0.3593
			(1e-04)	(1e-04)	(7e-04)	(3e-04)	(5e-04)	(0.0038)
0.25	25	5	0.0498	0.0847	0.1149	0.2464	0.4258	0.5614
			(6e-04)	(0.0012)	(0.0021)	(0.0017)	(0.0046)	(0.0082)
0.25	50	20	0.0148	0.0177	0.0277	0.0779	0.0947	0.1462
			(1e-04)	(1e-04)	(2e-04)	(2e-04)	(5e-04)	(9e-04)
0.5	25	5	0.0437	0.084	0.0934	0.2228	0.4493	0.4636
			(5e-04)	(0.0011)	(0.0016)	(0.0014)	(0.0046)	(0.0065)
0.5	50	20	0.0133	0.0241	0.021	0.0703	0.129	0.1104
			(1e-04)	(2e-04)	(2e-04)	(2e-04)	(7e-04)	(5e-04)
0.75	25	5	0.0502	0.1122	0.116	0.2477	0.5954	0.5689
			(6e-04)	(0.0017)	(0.0023)	(0.0017)	(0.0064)	(0.0089)
0.75	50	20	0.0148	0.0407	0.0282	0.0779	0.2101	0.1452
			(1e-04)	(4e-04)	(2e-04)	(2e-04)	(0.0014)	(8e-04)
0.9	25	5	0.0642	0.1778	0.1947	0.3052	0.9068	0.9579
			(9e-04)	(0.0031)	(0.0046)	(0.0024)	(0.0125)	(0.021)
0.9	50	20	0.0182	0.0642	0.0596	0.0954	0.3349	0.3554
			(1e-04)	(6e-04)	(7e-04)	(3e-04)	(0.0024)	(0.0034)

Table 5.3
Coverage probabilities of 95% credible interval for β.

p	n	T	Design I	Design II	Design III	Design IV	Design V	Design VI
0.1	25	5	0.93	0.975	0.905	0.96	0.995	0.735
0.1	50	20	0.97	0.95	0.915	0.9	0.9	0.775
0.25	25	5	0.935	0.95	0.935	0.97	0.975	0.935
0.25	50	20	0.955	0.965	0.925	0.915	0.895	0.89
0.5	25	5	0.93	0.935	0.97	0.985	0.915	0.975
0.5	50	20	0.95	0.925	0.94	0.93	0.84	0.965
0.75	25	5	0.95	0.93	0.975	0.965	0.795	0.93
0.75	50	20	0.96	0.93	0.95	0.925	0.81	0.91
0.9	25	5	0.95	0.92	0.945	0.965	0.75	0.74
0.9	50	20	0.97	0.93	0.91	0.92	0.83	0.8

6. Discussion and future works

In this thesis, firstly we propose a new method for nonparametric Bayesian quantile regression based on the Dirichlet process mixture of logistic distributions, where the mixture is taken over both the location parameter and scale parameter. We carefully study how the constraint impacts the inference of the regression parameters and develop a simple adjustment to get correct inference of the regression coefficients even when the quantile constraint is violated. And we are able to show that our proposed model is equivalent to the model which employs location shift of the mixture to satisfy the quantile constraint. We thus avoid the usual complication in constructing a mixture kernel density to satisfy the quantile constraint. As a result, the proposed model has a simpler kernel and is yet flexible. Secondly, we provide theoretical guarantee on the posterior consistency of our proposed model. Thirdly we propose a modification to handle data with heteroscedasticity. Efficient MCMC algorithms for the posterior inference are also provided. Simulation studies show that our method works as well as the DPMMN method in terms of accuracy, while our method is faster in computation and more robust to outliers. Fourthly, we propose a model to handle longitudinal data and the performance is evaluated by simulation study. We now summarize some directions of future works.

1. A direct extension of this thesis is to apply our approach to the nonparametric Bayesian mean regression. DPM-based mean regression usually requires the error density to be symmetric, which is sufficient for the mean constraint but is not necessary. And this symmetry requirement is very restrictive. Instead, given any

probability measure P over $\mathbb{R} \times \mathbb{R}^+$ and any location-scale family k with finite moment, define

$$\mu_P := \int x \int \frac{1}{\sigma} k\left(\frac{x-\mu}{\sigma}\right) dP(\mu, \sigma) dx.$$

We can define a probability measure over the space of densities with mean 0, by $f(x) = \int \frac{1}{\sigma} k\left(\frac{x-\mu-\mu_P}{\sigma}\right) dP(\mu, \sigma)$ with $P \sim DP(\alpha, G)$. This is a more flexible model. And the arguments in Section 2.2 holds for any constraint on the location parameter. Thus, for the mean regression, we can just model the error distribution by a location-scale DPM of normal densities and adjust the inference for the intercept properly. However, in the mean regression case, more effort is required to prove the posterior consistency. The main barrier is to derive a variant of Swartz's theorem for this case. As for the tail behaviour of μ_P, a simple application of the Sethuraman construction [95] gives us $E(\mu_P) = \int \mu dG(\mu, \sigma)$, thus a bound for the tail probability can be derived by Markov inequality. If a variant of Swartz's theorem can be derived, all the other arguments in this section can be translated to the mean regression case with no difficulty.

2. We have not studied the posterior consistency of our proposed model in Chapter 5 for longitudinal data. When there is no random effects, the posterior consistency can be derived following the approaches in Chapter 3. However, in the presence of random effects, substantial modification is required to show the posterior consistency. We also need to impose constraints on the ratio T/n to guarantee the posterior consistency. The key is to extend Schwartz's theorem to the scenario that the number of regression coefficients goes to infinity as the sample size grows. We will work on this problem in future works. Note that in the frequentist literature, e.g. [58], to guarantee the posterior consistency, one has to require $n/T^s \to 0$ as

$T, n \to \infty$ for $s \geq 1$. This is a stringent condition on the sample size growth rate. It would be nice to obtain posterior consistency for the Bayesian approach with less restrictive conditions on the sample growth rate.

3. There are still many gaps to be filled in the theory on the rate of convergence for nonparametric Bayesian methods. Recently [98] showed that, for the BALD method, the Bayes estimates for the regression coefficients are still consistent and attain certain rate of convergence when the ALD is misspecified, under the condition that the tail of the true error distribution is not too heavy along with other reasonable conditions. This result justifies the application of BALD for quantile regression for most of cases except for the heavy-tailed error distribution like Cauchy. On the other hand, for nonparametric Bayesian especially DPM models, most of the results focus on the consistency or convergence rates for the density estimation [37,39,40,42,97,108,110]. In the regression context, [41] derived the posterior rate of convergence for the joint estimates for the error density and regression coefficients. Under stringent conditions, the obtained convergence rate is slower than $n^{-1/2}$. And the derived convergence rate is only a lower bound, so we can not actually use these rates to compare the efficiency of two nonparametric Bayesian models. In the regression context, the convergence rate of the parametric parts (regression coefficients) is more important. Even though jointly, the parametric part and nonparametric part have a slow convergence rate, the parametric part alone may still have a root-n rate. It is desirable to show that applying nonparametric Bayesian methods for regression can attain a root-n rate for the regression coefficients, thus it is superior to the simple BALD. [8, 96] established conditions to guarantee the asymptotic normality of the marginal posterior for the parametric parts in a gen-

eral semiparametric estimation problem. However, it is non-trivial to verify those conditions for a specific problem. One direction of future work is to establish those conditions for the nonparametric Bayesian regression problems.

4. A closely related topic to this thesis is simultaneous quantile regression. The goal is either to get regression coefficients of multiple quantiles simultaneously or estimate the regression coefficients as functions of quantile. One main difficulty is to guarantee the monotonicity of the quantile function. For the frequentist methods, multiple quantiles can be treated simultaneously by adding a monotonicity constraint to the optimization problem for example as in [11, 73]. However, for the Bayesian approach, unless to model the response variable and the covariates jointly as in [99], one has to add this monotonicity constraint in a more elaborate way. For instance, in [103], the authors presented a simpler equivalent characterization of the monotonicity constraint through an interpolation of two monotone curves which are modelled via logistic transformations of a smooth Gaussian process. [88, 90, 91] defined the quantile process as the linear combination of some basis functions, and the monotonicity is guaranteed by putting constraints on the coefficients of the linear combination. A further extension of this thesis is to generalize our proposed models to handle simultaneous quantile regression.

5. Another possible extension of this thesis is to also model the regression function in a nonparametric way. One may either use cubic spline as in [22, 101] or put a Gaussian process prior to the regression function as in [12].

6. When there are multiple nonparametric Bayesian models for the same data set, one should be able to compare different models and choose the best one. This

is usually done through the Bayes factor (see [57]) or marginal likelihood which can be calculated by the method of [21] and extended by [7]. Usually, to perform such calculations in nonparametric Bayesian problems, one requires the conjugacy between the kernel density and the base measure. However, in the current literature, there is no method to calculate the marginal likelihood for the non-conjugate case. This may also be an interesting future work.

7. It is desirable to perform quantile regression analysis when the response variable is multivariate in some applications, for example, the multivariate growth charts [112]. However, for multivariate distributions, the quantile function is not uniquely defined, since there is no inherent ordering in multi-dimension. Based on different ways to order multivariate observations, there are multiple definitions for multivariate quantile functions [14–17, 20, 84]. Frequentist multivariate quantile regression still relies on minimizing a loss function directly related to the definition of multivariate quantile [18, 50, 112]. [28] proposed a Bayesian multivariate quantile regression methods based on the multivariate substitution likelihood, which generalizes the univariate version [69]. [107] developed a Bayesian bivariate quantile regression model by using a multivariate version of the location scale mixture representation for the asymmetric Laplace distribution. A possible extension of this thesis is to apply the DPM models for multivariate quantile regression. A simple idea is to use the multivariate normal distribution as the kernel and propose some adjustments to guarantee that the mixture satisfies a version of quantile constraint.

7. Appendix

A.1. Proof of Proposition 2.2.1

Proof Let $E^{(M)}(\cdot|\mathbf{x},\mathbf{y})$, $Var^{(M)}(\cdot|\mathbf{x},\mathbf{y})$ and $Cov^{(M)}(\cdot,\cdot|\mathbf{x},\mathbf{y})$ denote the posterior mean, variance and covariance under model (M), respectively.

To simplify the equations, let $\hat{\beta}_0^{\,(A)}$ and $\hat{\beta}_0^{\,(B)}$ denote the posterior mean of β_0 in (A) and (B) respectively. Let $\hat{V}(\beta_0)^{(B)}$ denote the posterior variance of β_0 in (B). Also let $\hat{\mu}_F^{(A)}$ denote the posterior mean of q_F in (A) and let $\hat{V}(\beta_0 - q_F)^{(A)}$ denote the posterior variance of $\beta_0 - q_F$ in (A). Let $d\Pi(\beta_0,\beta_1,F)$ denote $d\beta_0 d\beta_1 d\Lambda(F)$ and let $d\Pi(\beta_0,\beta_1,F^*)$ denote $d\beta_0 d\beta_1 d\Lambda^*(F^*)$. Let f_i denote $f_f(y_i-\beta_0-\beta_1 x_i)$ and let f_i^* denote $f_{F^*}(y_i-\beta_0-\beta_1 x_i)$. Also let f_{i,q_F} denote $f_F(y_i-\beta_0-\beta_1 x_i-q_F)$. Note that $f_{i,q_F}=f_i^*$ by definition.

In model (A), if the prior of β_0 is $\pi_1(\beta_0)\propto 1$, the posterior distribution of β_1 is given by

$$\begin{aligned}
\beta_1|y_1,\cdots,y_n &\propto \pi_2(\beta_1)\int\prod_{i=1}^n f_i\pi_1(\beta_0)d\Lambda(F)d\beta_0\\
&= \pi_2(\beta_1)\int\prod_{i=1}^n f_{i,q_F}\pi_1(\beta_0+q_F)d\beta_0 d\Lambda(F)\\
&= \pi_2(\beta_1)\int\prod_{i=1}^n f_i^*\pi_1(\beta_0)d\beta_0 d\Lambda^*(F^*),
\end{aligned}$$

which is the posterior distribution of β_1 in model (B).

As for the relation between the posterior means and variance of β_0 in model (A) and (B).

$$
\begin{aligned}
& E^{(A)}(\beta_0|\mathbf{x},\mathbf{y}) \\
=& \frac{\int \beta_0 \pi_1(\beta_0) \prod_{i=1}^{n} f_i \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F)}{\int \pi_1(\beta_0) \prod_{i=1}^{n} f_i \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F)} \\
=& \frac{\int (\beta_0+q_F) \pi_1(\beta_0+q_F) \prod_{i=1}^{n} f_{i,q_F} \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F)}{\int \pi_1(\beta_0+q_F) \prod_{i=1}^{n} f_{i,q_F} \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F)} \\
=& \frac{\int \beta_0 \pi_1(\beta_0) \prod_{i=1}^{n} f_{i,q_F} \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F)}{\int \pi_1(\beta_0) \prod_{i=1}^{n} f_{i,q_F} \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F)} \\
& + \frac{\int q_F \pi_1(\beta_0) \prod_{i=1}^{n} f_{i,q_F} \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F)}{\int \pi_1(\beta_0) \prod_{i=1}^{n} f_{i,q_F} \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F)} \\
=& \frac{\int \beta_0 \pi_1(\beta_0) \prod_{i=1}^{n} f_i^* \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F^*)}{\int \pi_1(\beta_0) \prod_{i=1}^{n} f_i^* \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F^*)} \\
& + \frac{\int q_F \pi_1(\beta_0) \prod_{i=1}^{n} f_i \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F)}{\int \pi_1(\beta_0) \prod_{i=1}^{n} f_i \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F)} \\
=& E^{(B)}(\beta_0|\mathbf{x},\mathbf{y}) + E^{(A)}(q_F|\mathbf{x},\mathbf{y}).
\end{aligned}
$$

$$
\begin{aligned}
&Var^{(A)}(\beta_0|\mathbf{x},\mathbf{y}) \\
=&\frac{\int \left(\beta_0 - \hat{\beta_0}^{(A)}\right)^2 \pi_1(\beta_0) \prod_{i=1}^{n} f_i \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F)}{\int \pi_1(\beta_0) \prod_{i=1}^{n} f_i \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F)} \\
=&\frac{\int \left(\beta_0 - \hat{\beta_0}^{(A)} + q_F\right)^2 \pi_1(\beta_0+q_F) \prod_{i=1}^{n} f_{i,q_F} \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F)}{\int \pi_1(\beta_0+q_F) \prod_{i=1}^{n} f_{i,q_F} \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F)} \\
=&\frac{\int \left(\beta_0 - \hat{\beta_0}^{(B)} - \hat{\mu}_F^{(A)} + q_F\right)^2 \pi_1(\beta_0+q_F) \prod_{i=1}^{n} f_{i,q_F} \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F)}{\int \pi_1(\beta_0+q_F) \prod_{i=1}^{n} f_{i,q_F} \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F)} \\
=&\frac{\int \left(\beta_0 - \hat{\beta_0}^{(B)}\right)^2 \pi_1(\beta_0) \prod_{i=1}^{n} f_{i,q_F} \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F)}{\int \pi_1(\beta_0) \prod_{i=1}^{n} f_{i,q_F} \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F)} \\
&+ 2 \cdot \frac{\int \left(\beta_0 - \hat{\beta_0}^{(A)} - (q_F - \hat{\mu}_F^{(A)})\right)\left(q_F - \hat{\mu}_F^{(A)}\right) \pi_1(\beta_0) \prod_{i=1}^{n} f_i \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F)}{\int \pi_1(\beta_0) \prod_{i=1}^{n} f_i \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F)} \\
&+ \frac{\int \left(q_F - \hat{\mu}_F^{(A)}\right)^2 \pi_1(\beta_0) \prod_{i=1}^{n} f_i \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F)}{\int \pi_1(\beta_0) \prod_{i=1}^{n} f_i \pi_2(\beta_1) d\Pi(\beta_0,\beta_1,F)} \\
=&Var^{(B)}(\beta_0|\mathbf{x},\mathbf{y}) + 2 \cdot Cov^{(A)}(\beta_0, q_F|\mathbf{x},\mathbf{y}) - Var^{(A)}(q_F|\mathbf{x},\mathbf{y}).
\end{aligned}
$$

The result follows by collecting terms and applying the formula $Var(X-Y) = Var(X) + Var(Y) - 2Cov(X,Y)$. ■

A.2. Proof of Proposition 4.1.1

Proof Let $E^{(M)}(\cdot|\mathbf{x},\mathbf{y})$, $Var^{(M)}(\cdot|\mathbf{x},\mathbf{y})$ and $Cov^{(M)}(\cdot,\cdot|\mathbf{x},\mathbf{y})$ denote the posterior mean, variance and covariance under model (M), respectively.

To simplify the equations, let $\hat{\beta_0}^{(A')}$, $\hat{\beta_0}^{(B')}$, $\hat{\beta_1}^{(A')}$ and $\hat{\beta_1}^{(B')}$ denote the posterior mean of β_0 and β_1 in (A') and (B') respectively. Let $\hat{V}(\beta_0)^{(B')}$ and $\hat{V}(\beta_1)^{(B')}$ denote the posterior variance of β_0 and β_1 in (B'). Also let $\hat{\mu}_F^{(A')}$ and $\hat{\gamma q}_F^{(A')}$ denote the posterior mean of q_F and γq_F in (A') and let $\hat{V}(\beta_0 - q_F)^{(A')}$ and $\hat{V}(\beta_1 - \gamma q_F)^{(A')}$ denote the posterior variance

of $\beta_0 - q_F$ in (A). Let $d\Pi(\beta_0, \beta_1, F) := d\beta_0 d\beta_1 d\Lambda(F)$, $d\Pi(\beta_0, \beta_1, F^*) := d\beta_0 d\beta_1 d\Lambda^*(F^*)$, $d\Pi(\beta_0, \beta_1, \gamma, F) := d\beta_0 d\beta_1 d\gamma d\Lambda(F)$ and $d\Pi(\beta_0, \beta_1, \gamma, F^*) := d\beta_0 d\beta_1 d\gamma d\Lambda^*(F^*)$. Let $f_i := f_F\left(\frac{y_i - \beta_0 - \beta_1 x_i}{1+\gamma x_i}\right)$, $f_{i,q_F} := f_F\left(\frac{y_i - \beta_0 - \beta_1 x_i}{1+\gamma x_i} - q_F\right)$. Also let $f_i^* := f_{F^*}\left(\frac{y_i - \beta_0 - \beta_1 x_i}{1+\gamma x_i}\right)$. Note that by definition $f_{i,q_F} = f_i^*$.

First we compare the posterior distributions of γ.

$$\text{Posterior distribution of } \gamma \text{ in } (B') \propto \int \prod_{i=1}^{n} f_i^* \pi_1(\beta_0)\pi_2(\beta_1) d\Pi(\beta_0, \beta_1, F^*).$$

If $\pi_1(\beta_0) \propto 1$ and $\pi_2(\beta_1) \propto 1$, we have

$$\begin{aligned}
&\text{Posterior distribution of } \gamma \text{ in } (A') \\
&\propto \int \prod_{i=1}^{n} f_i \pi_1(\beta_0)\pi_2(\beta_1) d\Pi(\beta_0, \beta_1, F) \\
&= \int \prod_{i=1}^{n} f_{i,q_F} \pi_1(\beta_0 + q_F)\pi_2(\beta_1 + \gamma q_F) d\Pi(\beta_0, \beta_1, F) \\
&= \int \prod_{i=1}^{n} f_i^* \pi_1(\beta_0)\pi_2(\beta_1) d\Pi(\beta_0, \beta_1, F^*).
\end{aligned}$$

So the posterior distributions of γ are the same in models (A') and (B'). We now compare the posterior means and variances of β_0, β_1 and γ for the two models.

If $\pi_1(\beta_0) \propto 1$ and $\pi_2(\beta_1) \propto 1$, we have

$$
\begin{aligned}
& E^{(A')}(\beta_0|\mathbf{x},\mathbf{y}) \\
=& \frac{\int \beta_0 \prod_{i=1}^{n} f_i \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)}{\int \prod_{i=1}^{n} f_i \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)} \\
=& \frac{\int \beta_0 \prod_{i=1}^{n} f_{i,q_F} \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)}{\int \prod_{i=1}^{n} f_{i,q_F} \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)} \\
=& \frac{\int (\beta_0+q_F) \prod_{i=1}^{n} f_{i,q_F} \pi_1(\beta_0+q_F)\pi_2(\beta_1+\gamma q_F)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)}{\int \prod_{i=1}^{n} f_{i,q_F} \pi_1(\beta_0+q_F)\pi_2(\beta_1+\gamma q_F)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)} \\
=& \frac{\int \beta_0 \prod_{i=1}^{n} f_i^* \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F^*)}{\int \prod_{i=1}^{n} f_i^* \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F^*)} \\
& + \frac{\int q_F \prod_{i=1}^{n} f_{i,q_F} \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)}{\int \prod_{i=1}^{n} f_{i,q_F} \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)} \\
=& E^{(B')}(\beta_0|\mathbf{x},\mathbf{y}) \\
& + \frac{\int q_F \prod_{i=1}^{n} f_i \pi_1(\beta_0-q_F)\pi_2(\beta_1-\gamma q_F)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)}{\int \prod_{i=1}^{n} f_i \pi_1(\beta_0-q_F)\pi_2(\beta_1-\gamma q_F)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)} \\
=& E^{(B')}(\beta_0|\mathbf{x},\mathbf{y}) + E^{(A')}(q_F|\mathbf{x},\mathbf{y}).
\end{aligned}
$$

$$
\begin{aligned}
&Var^{(A')}(\beta_0|\mathbf{x},\mathbf{y}) \\
=&\frac{\int \left(\beta_0-\hat{\beta}_0^{(A')}\right)^2 \prod_{i=1}^{n} f_i\pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma)d\Pi(\beta_0,\beta_1,\gamma,F)}{\int \prod_{i=1}^{n} f_i\pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma)d\Pi(\beta_0,\beta_1,\gamma,F)} \\
=&\frac{\int \left(\beta_0-\hat{\beta}_0^{(A')}\right)^2 \prod_{i=1}^{n} f_{i,q_F}\pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma)d\Pi(\beta_0,\beta_1,\gamma,F)}{\int \prod_{i=1}^{n} f_{i,q_F}\pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma)d\Pi(\beta_0,\beta_1,\gamma,F)} \\
=&\frac{\int \left(\beta_0-\hat{\beta}_0^{(A')}+q_F\right)^2 \prod_{i=1}^{n} f_{i,q_F}\pi_1(\beta_0+q_F)\pi_2(\beta_1+\gamma q_F)\pi_4(\gamma)d\Pi(\beta_0,\beta_1,\gamma,F)}{\int \prod_{i=1}^{n} f_{i,q_F}\pi_1(\beta_0+q_F)\pi_2(\beta_1+\gamma q_F)\pi_4(\gamma)d\Pi(\beta_0,\beta_1,\gamma,F)} \\
=&\frac{\int \left(\beta_0-\hat{\beta}_0^{(B')}+q_F-\hat{\mu}_F^{(A')}\right)^2 \prod_{i=1}^{n} f_{i,q_F}\pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma)d\Pi(\beta_0,\beta_1,\gamma,F)}{\int \prod_{i=1}^{n} f_{i,q_F}\pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma)d\Pi(\beta_0,\beta_1,\gamma,F)} \\
=&\frac{\int \left(\beta_0-\hat{\beta}_0^{(B')}\right)^2 \prod_{i=1}^{n} f_i^*\pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma)d\Pi(\beta_0,\beta_1,\gamma,F^*)}{\int \prod_{i=1}^{n} f_i^*\pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma)d\Pi(\beta_0,\beta_1,\gamma,F^*)} \\
&+2\cdot\frac{\int \left(\beta_0-\hat{\beta}_0^{(A')}\right)\left(q_F-\hat{\mu}_F^{(A')}\right) \prod_{i=1}^{n} f_i\pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma)d\Pi(\beta_0,\beta_1,\gamma,F)}{\int \prod_{i=1}^{n} f_i\pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma)d\Pi(\beta_0,\beta_1,\gamma,F)} \\
&-\frac{\int \left(q_F-\hat{\mu}_F^{(A')}\right)^2 \prod_{i=1}^{n} f_i\pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma)d\Pi(\beta_0,\beta_1,\gamma,F)}{\int \prod_{i=1}^{n} f_i\pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma)d\Pi(\beta_0,\beta_1,\gamma,F)} \\
=&Var^{(B')}(\beta_0|\mathbf{x},\mathbf{y})+2Cov^{(A')}(\beta_0,q_F|\mathbf{x},\mathbf{y})-Var^{(A')}(q_F|\mathbf{x},\mathbf{y}).
\end{aligned}
$$

If $\pi_1(\beta_0) \propto 1$ and $\pi_2(\beta_1) \propto 1$, we have

$$
\begin{aligned}
& E^{(A')}(\beta_1|\mathbf{x},\mathbf{y}) \\
=& \frac{\int \beta_1 \prod_{i=1}^{n} f_i \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)}{\int \prod_{i=1}^{n} f_i \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)} \\
=& \frac{\int \beta_1 \prod_{i=1}^{n} f_{i,q_F} \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)}{\int \prod_{i=1}^{n} f_{i,q_F} \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)} \\
=& \frac{\int (\beta_1+\gamma q_F) \prod_{i=1}^{n} f_{i,q_F} \pi_1(\beta_0+q_F)\pi_2(\beta_1+\gamma q_F)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)}{\int \prod_{i=1}^{n} f_{i,q_F} \pi_1(\beta_0+q_F)\pi_2(\beta_1+\gamma q_F)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)} \\
=& \frac{\int \beta_1 \prod_{i=1}^{n} f_i^* \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F^*)}{\int \prod_{i=1}^{n} f_i^* \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F^*)} \\
& + \frac{\int \gamma q_F \prod_{i=1}^{n} f_{i,q_F} \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)}{\int \prod_{i=1}^{n} f_{i,q_F} \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)} \\
=& E^{(B')}(\beta_1|\mathbf{x},\mathbf{y}) \\
& + \frac{\int \gamma q_F \prod_{i=1}^{n} f_i \pi_1(\beta_0-q_F)\pi_2(\beta_1-\gamma q_F)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)}{\int \prod_{i=1}^{n} f_i \pi_1(\beta_0-q_F)\pi_2(\beta_1-\gamma q_F)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)} \\
=& E^{(B')}(\beta_1|\mathbf{x},\mathbf{y}) + E^{(A')}(\gamma q_F|\mathbf{x},\mathbf{y}).
\end{aligned}
$$

$$
\begin{aligned}
&Var^{(A')}(\beta_1|\mathbf{x},\mathbf{y}) \\
=&\frac{\int \left(\beta_1 - \hat{\beta}_1^{(A')}\right)^2 \prod_{i=1}^{n} f_i \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)}{\int \prod_{i=1}^{n} f_i \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)} \\
=&\frac{\int \left(\beta_1 - \hat{\beta}_1^{(A')}\right)^2 \prod_{i=1}^{n} f_{i,q_F} \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)}{\int \prod_{i=1}^{n} f_{i,q_F} \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)} \\
=&\frac{\int \left(\beta_1 - \hat{\beta}_1^{(A')} + \gamma q_F\right)^2 \prod_{i=1}^{n} f_{i,q_F} \pi_1(\beta_0+q_F)\pi_2(\beta_1+\gamma q_F)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)}{\int \prod_{i=1}^{n} f_{i,q_F} \pi_1(\beta_0+q_F)\pi_2(\beta_1+\gamma q_F)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)} \\
=&\frac{\int \left(\beta_1 - \hat{\beta}_1^{(B')} + \gamma q_F - \gamma \hat{q}_F{}^{(A')}\right)^2 \prod_{i=1}^{n} f_{i,q_F} \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)}{\int \prod_{i=1}^{n} f_{i,q_F} \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)} \\
=&\frac{\int \left(\beta_1 - \hat{\beta}_1^{(B')}\right)^2 \prod_{i=1}^{n} f_i^* \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F^*)}{\int \prod_{i=1}^{n} f_i^* \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F^*)} \\
&+ 2 \cdot \frac{\int \left(\beta_1 - \hat{\beta}_1^{(B')}\right)\left(\gamma q_F - \gamma \hat{q}_F{}^{(A')}\right) \prod_{i=1}^{n} f_i \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)}{\int \prod_{i=1}^{n} f_i \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)} \\
&- \frac{\int \left(\gamma q_F - \gamma \hat{q}_F{}^{(A')}\right)^2 \prod_{i=1}^{n} f_i \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)}{\int \prod_{i=1}^{n} f_i \pi_1(\beta_0)\pi_2(\beta_1)\pi_4(\gamma) d\Pi(\beta_0,\beta_1,\gamma,F)} \\
=&Var^{(B')}(\beta_1|\mathbf{x},\mathbf{y}) + 2Cov^{(A')}(\beta_1,\gamma q_F|\mathbf{x},\mathbf{y}) - Var^{(A')}(\gamma q_F|\mathbf{x},\mathbf{y}).
\end{aligned}
$$

The result follows by collecting terms and applying the formula $Var(X-Y) = Var(X) + Var(Y) - 2Cov(X,Y)$. ■

A.3. Additional tables

Table 7.1
MSE of regression coefficients in Design 1. MSE is reported as 100 × average (100 × standard error) over the 200 simulated data sets

p	Coef	FQR	BASL	DPMU	DPMMN	DPML	DPMUH	DPMMNH	DPMLH
0.5	β_0	1.558	1.385	1.481	1.072	1.145	1.440	1.063	1.142
		(0.163)	(0.137)	(0.151)	(0.105)	(0.110)	(0.147)	(0.103)	(0.110)
0.5	β_1	1.681	1.359	1.418	1.046	1.043	1.282	1.099	1.086
		(0.157)	(0.128)	(0.130)	(0.098)	(0.096)	(0.116)	(0.100)	(0.098)
0.5	β_2	1.482	1.261	1.476	1.069	1.082	1.380	1.151	1.162
		(0.149)	(0.136)	(0.179)	(0.115)	(0.116)	(0.149)	(0.121)	(0.121)
0.9	β_0	3.122	2.761	5.037	2.472	2.155	4.829	2.811	2.252
		(0.300)	(0.270)	(0.448)	(0.234)	(0.218)	(0.438)	(0.268)	(0.227)
0.9	β_1	2.947	2.403	1.487	1.061	1.055	2.242	2.130	1.942
		(0.303)	(0.239)	(0.140)	(0.097)	(0.096)	(0.233)	(0.211)	(0.187)
0.9	β_2	3.749	3.114	1.555	1.073	1.074	2.598	2.422	2.195
		(0.477)	(0.383)	(0.193)	(0.115)	(0.115)	(0.311)	(0.273)	(0.246)

Table 7.2
Lengths of 90% credible or confidence intervals and coverage probabilities (CP) for regression coefficients in Design 1.

p	Coef		FQR	BASL	DPMU	DPMMN	DPML	DPMUH	DPMMNH	DPMLH
0.5	β_0	CP	0.91	0.83	0.925	0.905	0.865	0.915	0.915	0.87
		Length	0.412	0.336	0.446	0.35	0.331	0.446	0.355	0.334
0.5	β_1	CP	0.88	0.82	0.785	0.885	0.865	0.85	0.9	0.895
		Length	0.43	0.341	0.324	0.346	0.338	0.337	0.346	0.339
0.5	β_2	CP	0.895	0.845	0.825	0.905	0.89	0.85	0.9	0.89
		Length	0.411	0.334	0.324	0.341	0.334	0.33	0.34	0.333
0.9	β_0	CP	0.865	0.685	0.61	0.915	0.86	0.64	0.905	0.855
		Length	0.571	0.327	0.431	0.533	0.474	0.442	0.555	0.487
0.9	β_1	CP	0.88	0.69	0.775	0.89	0.855	0.765	0.9	0.905
		Length	0.561	0.328	0.321	0.346	0.339	0.411	0.512	0.478
0.9	β_2	CP	0.875	0.65	0.84	0.9	0.895	0.81	0.885	0.895
		Length	0.57	0.33	0.32	0.343	0.335	0.41	0.509	0.474

Table 7.3
MSE of regression coefficients in Design 2. MSE is reported as 100 × average (100 × standard error) over the 200 simulated data sets

p	Coef	FQR	BASL	DPMU	DPMMN	DPML	DPMUH	DPMMNH	DPMLH
0.5	β_0	3.034	2.681	2.779	2.865	2.487	2.757	2.871	2.552
		(0.276)	(0.249)	(0.26)	(0.268)	(0.236)	(0.25)	(0.269)	(0.238)
0.5	β_1	2.42	2.062	2.129	1.705	1.612	2.144	1.871	1.872
		(0.269)	(0.222)	(0.201)	(0.151)	(0.142)	(0.186)	(0.178)	(0.178)
0.5	β_2	2.489	2.234	2.227	1.99	1.836	2.219	2.108	2.094
		(0.229)	(0.215)	(0.239)	(0.197)	(0.188)	(0.246)	(0.206)	(0.206)
0.9	β_0	32.023	26.561	132.458	23.345	30.347	127.611	24.049	29.235
		(3.119)	(2.586)	(6.037)	(2.393)	(2.834)	(5.907)	(2.371)	(2.741)
0.9	β_1	30.775	22.838	2.788	1.574	1.662	8.075	10.379	9.188
		(2.388)	(1.751)	(0.306)	(0.135)	(0.151)	(0.84)	(1.1)	(0.999)
0.9	β_2	23.695	18.811	3.45	1.762	1.858	6.735	9.413	8.031
		(2.33)	(1.811)	(0.382)	(0.188)	(0.193)	(0.676)	(0.937)	(0.746)

Table 7.4
Lengths of 90% credible or confidence intervals and coverage probabilities (CP) for regression coefficients in Design 2.

p	Coef		FQR	BASL	DPMU	DPMMN	DPML	DPMUH	DPMMNH	DPMLH
0.5	β_0	CP	0.87	0.85	0.895	0.875	0.81	0.89	0.865	0.815
		Length	0.51	0.465	0.56	0.506	0.426	0.568	0.516	0.434
0.5	β_1	CP	0.905	0.905	0.835	0.935	0.92	0.89	0.915	0.915
		Length	0.536	0.476	0.425	0.471	0.451	0.468	0.479	0.463
0.5	β_2	CP	0.875	0.875	0.835	0.88	0.895	0.87	0.88	0.865
		Length	0.551	0.475	0.424	0.466	0.446	0.452	0.479	0.463
0.9	β_0	CP	0.865	0.565	0.125	0.895	0.635	0.13	0.895	0.67
		Length	1.692	0.825	0.723	1.614	1.091	0.753	1.692	1.148
0.9	β_1	CP	0.875	0.53	0.77	0.94	0.925	0.735	0.885	0.88
		Length	1.511	0.782	0.421	0.446	0.459	0.659	1.096	0.936
0.9	β_2	CP	0.895	0.6	0.72	0.91	0.895	0.75	0.91	0.85
		Length	1.544	0.782	0.421	0.444	0.454	0.644	1.095	0.942

Table 7.5
MSE of regression coefficients in Design 3. MSE is reported as 100 × average (100 × standard error) over the 200 simulated data sets

p	Coef	FQR	BASL	DPMU	DPMMN	DPML	DPMUH	DPMMNH	DPMLH
0.5	β_0	1.463	1.389	1.447	1.521	1.548	1.479	1.531	1.553
		(0.14)	(0.131)	(0.134)	(0.143)	(0.148)	(0.132)	(0.146)	(0.147)
0.5	β_1	1.647	1.493	2.138	1.535	1.612	1.912	1.602	1.653
		(0.199)	(0.189)	(0.262)	(0.194)	(0.206)	(0.24)	(0.205)	(0.213)
0.5	β_2	1.583	1.485	2.153	1.596	1.729	2.236	1.756	1.878
		(0.189)	(0.173)	(0.259)	(0.194)	(0.213)	(0.272)	(0.221)	(0.24)
0.9	β_0	8.51	7.676	22.488	7.248	4.97	21.288	8.491	5.208
		(0.936)	(0.811)	(1.526)	(0.732)	(0.518)	(1.491)	(0.825)	(0.525)
0.9	β_1	7.883	6.515	2.441	1.661	1.619	4.059	4.661	4.098
		(0.793)	(0.668)	(0.288)	(0.206)	(0.208)	(0.372)	(0.462)	(0.424)
0.9	β_2	9.795	7.93	2.819	1.688	1.777	5.347	5.62	5.129
		(0.968)	(0.899)	(0.3)	(0.205)	(0.217)	(0.657)	(0.598)	(0.583)

Table 7.6
Lengths of 90% credible or confidence intervals and coverage probabilities (CP) for regression coefficients in Design 3.

p	Coef		FQR	BASL	DPMU	DPMMN	DPML	DPMUH	DPMMNH	DPMLH
0.5	β_0	CP	0.855	0.87	0.925	0.875	0.875	0.92	0.89	0.875
		Length	0.376	0.362	0.431	0.387	0.379	0.436	0.395	0.389
0.5	β_1	CP	0.88	0.89	0.84	0.9	0.9	0.83	0.875	0.89
		Length	0.402	0.37	0.388	0.391	0.396	0.383	0.385	0.391
0.5	β_2	CP	0.88	0.885	0.83	0.9	0.89	0.815	0.855	0.86
		Length	0.4	0.371	0.388	0.393	0.398	0.382	0.386	0.393
0.9	β_0	CP	0.905	0.665	0.335	0.94	0.87	0.385	0.945	0.88
		Length	1.028	0.531	0.539	0.943	0.69	0.558	0.974	0.712
0.9	β_1	CP	0.87	0.645	0.79	0.905	0.915	0.755	0.89	0.86
		Length	0.943	0.517	0.373	0.396	0.404	0.487	0.721	0.629
0.9	β_2	CP	0.875	0.66	0.71	0.885	0.89	0.69	0.855	0.86
		Length	0.903	0.515	0.389	0.398	0.407	0.499	0.719	0.627

Table 7.7
MSE of regression coefficients in Design 4. MSE is reported as 100 × average (100 × standard error) over the 200 simulated data sets

p	Coef	FQR	BASL	DPMU	DPMMN	DPML	DPMUH	DPMMNH	DPMLH
0.5	β_0	2.619	2.327	4.598	2.397	2.844	2.135	1.647	1.64
		(0.278)	(0.252)	(0.441)	(0.26)	(0.31)	(0.195)	(0.156)	(0.157)
0.5	β_1	4.484	4.4	18.848	5.979	9.071	3.008	2.391	2.332
		(0.461)	(0.459)	(1.545)	(0.634)	(0.93)	(0.281)	(0.232)	(0.225)
0.9	β_0	4.382	3.838	14.157	7.427	4.222	4.462	3.557	2.738
		(0.482)	(0.406)	(1.121)	(0.865)	(0.51)	(0.514)	(0.409)	(0.341)
0.9	β_1	9.224	8.264	124.851	176.466	160.425	6.105	4.278	3.871
		(0.988)	(0.887)	(6.848)	(6.978)	(5.706)	(0.65)	(0.437)	(0.397)

Table 7.8
Lengths of 90% credible or confidence intervals and coverage probabilities (CP) for regression coefficients in Design 4.

p	Coef		FQR	BASL	DPMU	DPMMN	DPML	DPMUH	DPMMNH	DPMLH
0.5	β_0	CP	0.865	0.785	0.66	0.79	0.74	0.88	0.875	0.845
		Length	0.447	0.37	0.448	0.389	0.395	0.478	0.412	0.388
0.5	β_1	CP	0.795	0.82	0.52	0.795	0.76	0.86	0.885	0.855
		Length	0.582	0.561	0.8	0.653	0.774	0.545	0.502	0.483
0.9	β_0	CP	0.83	0.63	0.36	0.895	0.87	0.745	0.94	0.93
		Length	0.608	0.356	0.437	0.857	0.633	0.499	0.631	0.559
0.9	β_1	CP	0.81	0.685	0.035	0.005	0.005	0.75	0.935	0.94
		Length	0.92	0.564	0.757	0.745	0.74	0.617	0.756	0.692

Table 7.9
MSE of regression coefficients in Design 5. MSE is reported as 100 × average (100 × standard error) over the 200 simulated data sets

p	Coef	FQR	BASL	DPMU	DPMMN	DPML	DPMUH	DPMMNH	DPMLH
0.5	β_0	2.464	2.111	2.356	1.738	1.815	2.465	1.86	1.944
		(0.236)	(0.208)	(0.247)	(0.175)	(0.186)	(0.248)	(0.179)	(0.189)
0.5	β_1	1.608	1.277	1.506	1.074	1.095	1.445	1.111	1.106
		(0.147)	(0.113)	(0.157)	(0.096)	(0.097)	(0.133)	(0.099)	(0.099)
0.5	β_2	1.794	1.559	1.559	1.229	1.244	1.514	1.291	1.321
		(0.229)	(0.174)	(0.157)	(0.123)	(0.127)	(0.155)	(0.14)	(0.143)
0.5	β_3	0.536	0.403	0.425	0.341	0.346	0.57	0.414	0.426
		(0.058)	(0.044)	(0.065)	(0.037)	(0.038)	(0.113)	(0.046)	(0.048)
0.9	β_0	4.313	3.576	5.826	2.77	2.699	7.716	3.447	3.73
		(0.431)	(0.346)	(0.534)	(0.274)	(0.27)	(0.673)	(0.338)	(0.369)
0.9	β_1	3.357	2.425	1.459	1.111	1.1	2.103	2.024	1.905
		(0.344)	(0.235)	(0.141)	(0.099)	(0.1)	(0.211)	(0.197)	(0.18)
0.9	β_2	3.453	2.654	1.658	1.271	1.272	2.206	2.226	2.028
		(0.308)	(0.252)	(0.17)	(0.131)	(0.135)	(0.257)	(0.256)	(0.228)
0.9	β_3	0.885	0.673	0.449	0.353	0.344	1.151	0.828	1.063
		(0.086)	(0.073)	(0.064)	(0.042)	(0.038)	(0.176)	(0.081)	(0.117)

Table 7.10

Lengths of 90% credible or confidence intervals and coverage probabilities (CP) for regression coefficients in Design 5.

p	Coef		FQR	BASL	DPMU	DPMMN	DPML	DPMUH	DPMMNH	DPMLH
0.5	β_0	CP	0.845	0.845	0.9	0.885	0.865	0.88	0.865	0.865
		Length	0.498	0.418	0.509	0.436	0.419	0.526	0.447	0.436
0.5	β_1	CP	0.92	0.845	0.815	0.91	0.895	0.84	0.885	0.89
		Length	0.405	0.335	0.333	0.345	0.341	0.344	0.345	0.339
0.5	β_2	CP	0.86	0.82	0.805	0.87	0.87	0.81	0.87	0.855
		Length	0.396	0.331	0.332	0.344	0.338	0.343	0.344	0.337
0.5	β_3	CP	0.89	0.85	0.89	0.895	0.88	0.88	0.905	0.905
		Length	0.265*	0.185	0.187	0.192	0.188	0.228	0.218	0.225
0.9	β_0	CP	0.86	0.7	0.63	0.94	0.88	0.64	0.93	0.865
		Length	0.707	0.415	0.515	0.596	0.54	0.594	0.679	0.623
0.9	β_1	CP	0.865	0.735	0.825	0.905	0.895	0.855	0.91	0.92
		Length	0.562	0.333	0.328	0.348	0.34	0.432	0.514	0.481
0.9	β_2	CP	0.87	0.69	0.805	0.87	0.86	0.865	0.9	0.89
		Length	0.603	0.334	0.323	0.346	0.337	0.444	0.512	0.479
0.9	β_3	CP	0.935	0.745	0.845	0.895	0.875	0.87	0.95	0.935
		Length	0.862*	0.19	0.184	0.192	0.187	0.293	0.317	0.328

*Since the corresponding covariate is generated from t_2, FQR produces confidence interval with infinite length for some simulated data sets. As a result, the average length of the confidence interval is infinite. For comparison, we use the median of the lengths of the frequentist confidence interval. For other methods, there is little difference between the mean and the median of the lengths of credible intervals.

Table 7.11
MSE of regression coefficients in Design 6. MSE is reported as 100 × average (100 × standard error) over the 200 simulated data sets

p	Coef	FQR	BASL	DPMU	DPMMN	DPML	DPMUH	DPMMNH	DPMLH
0.5	β_0	2.074	1.698	1.712	1.461	1.464	1.68	1.446	1.466
		(0.2)	(0.17)	(0.174)	(0.148)	(0.151)	(0.171)	(0.147)	(0.151)
0.5	β_1	2.413	2.015	2.59	1.907	1.884	2.498	1.982	1.928
		(0.263)	(0.23)	(0.318)	(0.222)	(0.215)	(0.323)	(0.22)	(0.215)
0.5	β_2	2.127	1.915	2.456	1.85	1.849	2.617	2.005	1.994
		(0.185)	(0.16)	(0.23)	(0.159)	(0.158)	(0.222)	(0.168)	(0.164)
0.9	β_0	8.573	8.396	12.972	9.962	5.279	11.502	11.448	5.822
		(0.86)	(0.859)	(1.054)	(0.978)	(0.484)	(0.973)	(1.122)	(0.551)
0.9	β_1	9.938	8.085	2.706	2.006	1.858	5.646	5.423	4.88
		(1.081)	(0.884)	(0.315)	(0.232)	(0.218)	(0.607)	(0.642)	(0.575)
0.9	β_2	9.781	8.091	2.851	1.93	1.839	6.21	5.863	5.149
		(0.984)	(0.776)	(0.254)	(0.17)	(0.158)	(0.557)	(0.558)	(0.494)

Table 7.12
Lengths of 90% credible or confidence intervals and coverage probabilities (CP) for regression coefficients in Design 6.

p	Coef		FQR	BASL	DPMU	DPMMN	DPML	DPMUH	DPMMNH	DPMLH
0.5	β_0	CP	0.89	0.88	0.935	0.915	0.885	0.93	0.92	0.895
		Length	0.436	0.409	0.49	0.426	0.41	0.5	0.435	0.418
0.5	β_1	CP	0.885	0.855	0.83	0.895	0.895	0.85	0.87	0.86
		Length	0.457	0.413	0.424	0.431	0.431	0.435	0.427	0.428
0.5	β_2	CP	0.85	0.83	0.8	0.87	0.885	0.8	0.84	0.855
		Length	0.449	0.411	0.407	0.426	0.425	0.422	0.421	0.421
0.9	β_0	CP	0.875	0.65	0.54	0.92	0.895	0.57	0.92	0.895
		Length	1.011	0.542	0.596	0.966	0.718	0.589	1	0.737
0.9	β_1	CP	0.88	0.655	0.805	0.875	0.895	0.735	0.885	0.865
		Length	0.936	0.537	0.428	0.435	0.437	0.557	0.75	0.666
0.9	β_2	CP	0.88	0.595	0.765	0.87	0.88	0.665	0.87	0.85
		Length	0.897	0.522	0.411	0.426	0.43	0.529	0.739	0.653

Table 7.13
MSE of regression coefficients in the case with 5% outliers with $n = 100$. MSE is reported as 100 × average (100 × standard error) over the 200 simulated data sets

p	Coef	FQR	BASL	DPMU	DPMMN	DPML	DPMUH	DPMMNH	DPMLH
0.5	β_0	2.315	2.067	2.043	1.519	1.736	2.016	1.548	1.771
		(0.222)	(0.195)	(0.197)	(0.14)	(0.161)	(0.196)	(0.152)	(0.166)
0.5	β_1	1.752	1.431	1.127	0.973	0.914	1.171	1.145	1.177
		(0.166)	(0.138)	(0.099)	(0.098)	(0.097)	(0.109)	(0.122)	(0.125)
0.5	β_2	1.968	1.688	1.391	1.143	1.139	1.421	1.411	1.416
		(0.193)	(0.158)	(0.145)	(0.101)	(0.112)	(0.128)	(0.126)	(0.127)
0.9	β_0	22.965	31.396	8.268	1224.271	28.801	8.277	821.377	28.926
		(2.132)	(2.035)	(0.797)	(26.339)	(1.24)	(0.816)	(35.522)	(1.294)
0.9	β_1	10.768	8.865	1.756	0.74	0.587	2.965	3.202	1.463
		(1.189)	(1.027)	(0.206)	(0.085)	(0.066)	(0.366)	(0.547)	(0.274)
0.9	β_2	10.52	9.199	1.722	0.787	0.645	3.495	2.789	1.48
		(1.302)	(1.157)	(0.176)	(0.088)	(0.077)	(0.343)	(0.353)	(0.161)

Table 7.14
Lengths of 90% credible or confidence intervals and coverage probabilities (CP) for regression coefficients in the case with 5% outliers with $n = 100$.

p	Coef		FQR	BASL	DPMU	DPMMN	DPML	DPMUH	DPMMNH	DPMLH
0.5	β_0	CP	0.86	0.79	0.88	0.885	0.805	0.905	0.895	0.81
		Length	0.431	0.383	0.451	0.376	0.344	0.459	0.386	0.351
0.5	β_1	CP	0.925	0.875	0.85	0.94	0.935	0.885	0.905	0.9
		Length	0.438	0.385	0.324	0.355	0.347	0.351	0.37	0.367
0.5	β_2	CP	0.89	0.865	0.83	0.925	0.905	0.855	0.885	0.89
		Length	0.446	0.393	0.33	0.362	0.352	0.354	0.376	0.371
0.9	β_0	CP	0.53	0.305	0.535	0.02	0.11	0.515	0.115	0.16
		Length	2.773	0.787	0.483	3.105	0.632	0.474	3.52	0.687
0.9	β_1	CP	0.9	0.815	0.775	0.93	0.94	0.775	0.98	0.965
		Length	1.259	0.717	0.317	0.307	0.28	0.424	0.772	0.465
0.9	β_2	CP	0.915	0.825	0.775	0.915	0.955	0.725	0.98	0.97
		Length	1.277	0.719	0.323	0.31	0.284	0.442	0.767	0.467

Table 7.15
MSE of regression coefficients in the case with 5% outliers with $n = 500$. MSE is reported as 100 × average (100 × standard error) over the 200 simulated data sets

p	Coef	FQR	BASL	DPMU	DPMMN	DPML	DPMUH	DPMMNH	DPMLH
0.5	β_0	0.763	0.751	0.736	0.623	0.604	0.745	0.631	0.598
		(0.077)	(0.074)	(0.068)	(0.055)	(0.054)	(0.072)	(0.057)	(0.054)
0.5	β_1	0.385	0.356	0.186	0.073	0.039	0.243	0.176	0.198
		(0.034)	(0.031)	(0.019)	(0.009)	(0.004)	(0.022)	(0.014)	(0.016)
0.5	β_2	0.369	0.349	0.195	0.067	0.039	0.266	0.199	0.208
		(0.037)	(0.035)	(0.021)	(0.007)	(0.004)	(0.027)	(0.02)	(0.022)
0.9	β_0	13.374	13.647	2.28	218.249	15.82	2.269	114.011	15.699
		(0.53)	(0.499)	(0.272)	(35.605)	(0.406)	(0.346)	(26.989)	(0.403)
0.9	β_1	1.949	1.847	0.33	0.06	0.037	0.378	0.11	0.116
		(0.177)	(0.173)	(0.036)	(0.006)	(0.004)	(0.038)	(0.012)	(0.013)
0.9	β_2	1.674	1.605	0.376	0.075	0.048	0.467	0.122	0.132
		(0.161)	(0.155)	(0.037)	(0.008)	(0.005)	(0.045)	(0.013)	(0.014)

Table 7.16
Lengths of 90% credible or confidence intervals and coverage probabilities (CP) for regression coefficients in the case with 5% outliers with $n = 500$.

p	Coef		FQR	BASL	DPMU	DPMMN	DPML	DPMUH	DPMMNH	DPMLH
0.5	β_0	CP	0.73	0.32	0.725	0.705	0.605	0.74	0.71	0.625
		Length	0.194	0.087	0.198	0.169	0.152	0.196	0.17	0.152
0.5	β_1	CP	0.88	0.5	0.89	0.985	0.915	0.85	0.93	0.895
		Length	0.193	0.087	0.13	0.109	0.069	0.144	0.146	0.146
0.5	β_2	CP	0.88	0.56	0.835	0.965	0.92	0.825	0.92	0.9
		Length	0.197	0.087	0.131	0.109	0.069	0.145	0.146	0.145
0.9	β_0	CP	0.005	0.005	0.54	0.005	0	0.515	0.01	0
		Length	0.489	0.17	0.198	0.781	0.259	0.192	0.614	0.258
0.9	β_1	CP	0.835	0.48	0.75	0.98	0.92	0.77	0.91	0.88
		Length	0.445	0.166	0.129	0.107	0.069	0.145	0.119	0.112
0.9	β_2	CP	0.91	0.48	0.68	0.95	0.905	0.685	0.915	0.895
		Length	0.456	0.169	0.128	0.108	0.069	0.146	0.119	0.112

Table 7.17
MSE of regression coefficients in the case with 10% outliers with $n = 100$. MSE is reported as 100 × average (100 × standard error) over the 200 simulated data sets

p	Coef	FQR	BASL	DPMU	DPMMN	DPML	DPMUH	DPMMNH	DPMLH
0.5	β_0	4.438	4.259	3.693	3.72	4.089	3.707	3.751	3.974
		(0.372)	(0.329)	(0.304)	(0.296)	(0.308)	(0.309)	(0.288)	(0.302)
0.5	β_1	2.375	2.009	1.22	0.61	0.196	1.537	1.463	1.301
		(0.281)	(0.244)	(0.161)	(0.091)	(0.033)	(0.175)	(0.181)	(0.17)
0.5	β_2	2.206	1.824	1.402	0.734	0.207	1.831	1.415	1.346
		(0.221)	(0.169)	(0.204)	(0.098)	(0.03)	(0.219)	(0.134)	(0.131)
0.9	β_0	789.049	763.998	13.374	1657.398	1912.788	13.143	1682.601	1915.062
		(19.976)	(11.91)	(0.966)	(15.948)	(11.07)	(0.92)	(14.92)	(8.926)
0.9	β_1	123.868	85.224	1.554	0.264	0.159	2.794	0.187	0.147
		(7.675)	(5.871)	(0.152)	(0.033)	(0.018)	(0.343)	(0.024)	(0.016)
0.9	β_2	133.165	98.016	1.937	0.308	0.191	2.949	0.195	0.167
		(7.953)	(6.422)	(0.191)	(0.036)	(0.02)	(0.301)	(0.023)	(0.017)

Table 7.18
Lengths of 90% credible or confidence intervals and coverage probabilities (CP) for regression coefficients in the case with 10% outliers with $n = 100$.

p	Coef		FQR	BASL	DPMU	DPMMN	DPML	DPMUH	DPMMNH	DPMLH
0.5	β_0	CP	0.665	0.62	0.755	0.65	0.57	0.76	0.665	0.6
		Length	0.441	0.43	0.486	0.406	0.36	0.491	0.412	0.363
0.5	β_1	CP	0.885	0.9	0.83	0.95	0.935	0.82	0.87	0.905
		Length	0.448	0.433	0.311	0.282	0.163	0.35	0.356	0.342
0.5	β_2	CP	0.87	0.88	0.86	0.935	0.935	0.83	0.87	0.855
		Length	0.457	0.433	0.324	0.282	0.163	0.358	0.355	0.342
0.9	β_0	CP	0.025	0	0.44	0	0	0.425	0	0
		Length	4.222	1.696	0.487	2.063	0.718	0.488	2.008	0.718
0.9	β_1	CP	0.885	0.45	0.81	0.985	0.935	0.79	1	0.95
		Length	2.877	1.342	0.334	0.227	0.151	0.429	0.289	0.164
0.9	β_2	CP	0.9	0.405	0.76	0.965	0.945	0.79	1	0.95
		Length	2.874	1.321	0.335	0.232	0.156	0.426	0.295	0.168

Table 7.19
MSE of regression coefficients in the case with 10% outliers with $n = 500$. MSE is reported as 100 × average (100 × standard error) over the 200 simulated data sets

p	Coef	FQR	BASL	DPMU	DPMMN	DPML	DPMUH	DPMMNH	DPMLH
0.5	β_0	2.324	2.304	2.223	2.235	2.042	2.191	2.197	2.019
		(0.119)	(0.118)	(0.118)	(0.103)	(0.092)	(0.123)	(0.101)	(0.092)
0.5	β_1	0.348	0.329	0.132	0.028	0.019	0.241	0.157	0.156
		(0.037)	(0.035)	(0.014)	(0.002)	(0.002)	(0.024)	(0.014)	(0.014)
0.5	β_2	0.378	0.354	0.141	0.037	0.021	0.319	0.208	0.204
		(0.04)	(0.038)	(0.015)	(0.009)	(0.002)	(0.034)	(0.023)	(0.023)
0.9	β_0	783.02	778.617	13.501	1601.753	2003.172	13.748	1682.684	2002.886
		(10.777)	(9.297)	(0.801)	(30.528)	(3.608)	(0.821)	(27.797)	(3.612)
0.9	β_1	72.358	68.568	0.404	0.033	0.022	0.424	0.026	0.021
		(4.166)	(4.051)	(0.043)	(0.004)	(0.002)	(0.043)	(0.003)	(0.002)
0.9	β_2	78.413	74.256	0.449	0.031	0.02	0.479	0.023	0.019
		(4.481)	(4.29)	(0.042)	(0.003)	(0.002)	(0.049)	(0.003)	(0.002)

Table 7.20
Lengths of 90% credible or confidence intervals and coverage probabilities (CP) for regression coefficients in the case with 10% outliers with $n = 500$.

p	Coef		FQR	BASL	DPMU	DPMMN	DPML	DPMUH	DPMMNH	DPMLH
0.5	β_0	CP	0.28	0.04	0.27	0.14	0.1	0.28	0.175	0.095
		Length	0.204	0.09	0.204	0.181	0.156	0.206	0.183	0.156
0.5	β_1	CP	0.915	0.565	0.865	1	0.91	0.835	0.935	0.96
		Length	0.204	0.089	0.119	0.084	0.047	0.141	0.146	0.147
0.5	β_2	CP	0.935	0.515	0.88	0.97	0.905	0.8	0.895	0.905
		Length	0.204	0.089	0.119	0.085	0.049	0.14	0.146	0.146
0.9	β_0	CP	0	0	0.065	0	0	0.07	0	0
		Length	3.958	0.483	0.163	1.71	0.308	0.169	1.601	0.309
0.9	β_1	CP	0.9	0.14	0.625	0.98	0.88	0.67	1	0.88
		Length	2.124	0.363	0.113	0.085	0.047	0.124	0.094	0.048
0.9	β_2	CP	0.875	0.13	0.58	0.975	0.915	0.575	0.995	0.925
		Length	2.018	0.362	0.115	0.084	0.047	0.123	0.094	0.048

REFERENCES

[1] A. Abadie, J. Angrist, and G. Imbens. Instrumental variables estimates of subsidized training on the quantile of trainee earnings. *Econometrica*, 70:91–117, 2002.

[2] J. Abreveya. The effects of demographics and maternal behavior on the distribution of birth outcomes. *Empir. Econ.*, 26:247–257, 2001.

[3] R. Alhamzawi, K. Yu, and J. Pan. Prior elicitation in Bayesian quantile regression for longitudinal data. *J. Biomet. Biostat.*, 2:115, 2011.

[4] M. Amewou-Atisso, S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Posterior consistency for semi-parametric regression problems. *Bernoulli*, 9:291–312, 2003.

[5] C. E. Antoniak. Mixtures of Dirichlet processes with applications to bayesian nonparametric problems. *Ann. Statist.*, 2:1152–1174, 1974.

[6] O. Arias, K. Hallock, and W. Sosa-Escudero. Indidividual heterogeneity in the returns to schooling: instrumental variables quantile regression using twins data. *Empir. Econ.*, 26:7–40, 2001.

[7] S. Basu and S. Chib. Marginal likelihood and bayes factors for Dirichlet process mixture models. *J. Amer. Statist. Assoc.*, 98:224–235, 2003.

[8] P. J. Bickel and B. J. K. Kleijn. The semiparametric Bernstein-Von Mises theorem. *Ann. Statist.*, 40:206–237, 2012.

[9] D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. *Ann. Statist.*, 1:353–355, 1973.

[10] D. M. Blei and M. I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Anal.*, 1:121–144, 2006.

[11] H. D. Bondell, B. J. Reich, and H. Wang. Non-crossing quantile regression curve estimation. *Biometrika*, 97:825–838, 2010.

[12] A. Boukouvalas, R. Barillec, and D. Cornford. Gaussian process quantile regression using expectation propagation. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012.

[13] T. S. Breusch and A. R. Pagan. A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47:1287–1294, 1979.

[14] B. M. Brown and T. P. Hettmansperger. Affine invariant rank methods in the bivariate location model. *J. Roy. Statist. Soc. Ser. B*, 49:301–310, 1987.

[15] B. M. Brown and T. P. Hettmansperger. An affine invariant bivariate versions of the sign test. *J. Roy. Statist. Soc. Ser. B*, 51:117–125, 1989.

[16] Y. Cai. Multivariate quantile function models. *Statist. Sinica*, 20:481–496, 2010.

[17] B. Chakraborty. On affine equivariant multivariate quantiles. *Ann. Inst. Statist. Math.*, 53:380–403, 2001.

[18] B. Chakraborty. On multivariate quantile regression. *J. Statist. Plann. Inference*, 110:109–132, 2003.

[19] J. Chang and J. W. Fisher III. Parallel sampling of DP mixture models using sub-cluster splits. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 620–628. Curran Associates, Inc., 2013.

[20] P. Chaudhuri. On a geometric notation of quantiles for multivariate data. *J. Amer. Statist. Assoc.*, 91:862–872, 1996.

[21] S. Chib. Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.*, 90:1313–1321, 1995.

[22] S. Chib and E. Greenberg. Additive cubic spline regression with Dirichlet process mixture errors. *J. Econometrics*, 156:322–336, 2010.

[23] D. B. Dahl. Model-based clustering for expression data via a Dirichlet process mixture model. In K. Do, P. Müller, and M. Vannucci, editors, *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press, 2006.

[24] P.J. Diggle, P.J. Heagerty, K. Y. Liang, and S.L. Zeger. *Analysis of longitudinal data*. Oxford, UK: Oxford University Press, 2002.

[25] H. Doss and T. Sellke. The tails of probabilities chosen from a Dirichlet prior. *Ann. Statist.*, 10:1302–1305, 1982.

[26] D. B. Dunson, , and J. A. Taylor. Approximate Bayesian inference for quantiles. *J. Nonparametr. Stat.*, 17:385–400, 2004.

[27] D. B. Dunson, M. Watson, , and J. A. Taylor. Bayesian latent variable models for median regression on multiple outcomes. *Biometrics*, 14:296–304, 2003.

[28] D. B. Dunson, M. Watson, and J. A. Taylor. Multivariate quantiles and multiple-output regression quantiles: from L1 optimization to halfspace depth. *Biometrics*, 59:296–304, 2003.

[29] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.*, 90:577–588, 1995.

[30] Y. Feng, Y. Chen, and He. X. Bayesian quantile regression with approximate likelihood. *accepted by Bernoulli*, 2015.

[31] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1:209–230, 1973.

[32] T. S. Ferguson. Prior distribution on spaces of probability measures. *Ann. Statist.*, 2:615–629, 1974.

[33] T. S. Ferguson. Bayesian density estimation by mixture of normal distributions. In H. Rizvi and J. Rustagi, editors, *Recent Advances in Statistics*, pages 287–302. Academic Press, New York, 1983.

[34] G. B. Folland. *Real analysis: modern techniques and their applications*. Wiley, 2nd edition, 1999.

[35] M. Geraci and M. Bottai. Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics*, 8:140–154, 2007.

[36] R. H. Gerlach, C. W. S. Chen, and N. Y. C. Chan. Bayesian time-varying quantile forecasting for value-at-risk in financial markets. *J. Bus. Econ. Statist.*, 29:481–492, 2011.

[37] S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Posterior consistency of Dirichlet mixture in density estimation. *Ann. Statist.*, 27:143–158, 1999.

[38] S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.*, 27:143–158, 1999.

[39] S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28:500–531, 2000.

[40] S. Ghosal and A. W. van der Vaart. Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29:1233–1263, 2001.

[41] S. Ghosal and A. W. van der Vaart. Convergence rates of posterior distributions for noniid observation. *Ann. Statist.*, 35:192–223, 2007.

[42] S. Ghosal and A. W. van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.*, 35:697–723, 2007.

[43] J. Ghosh and S. Tokdar. Convergence and consistency of Newtons algorithm for estimating a mixing distribution. In J. Fan and H. Koul, editors, *The Frontiers of Statistics*, pages 429–443. London: Imperial College Press, 2006.

[44] J. K. Ghosh and R. V. Ramamoorthi. *Bayesian nonparametrics*. Springer, 2003.

[45] W. Gilks. Derivative-free adaptive rejection sampling for Gibbs sampling. In J. Bernardo, J. Berger, A. Dawid, and A. Smith, editors, *Bayesian Statistics 4*, pages 641–649. Oxford: Oxford University Press, 1992.

[46] W. Gilks and P. Wild. Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.*, 41:337–348, 1992.

[47] H. Glejser. A new test for heteroskedasticity. *J. Amer. Statist. Assoc.*, 64:315–323, 1969.

[48] Stephen M. Goldfeld and R. E. Quandt. Some tests for homoscedasticity. *J. Amer. Statist. Assoc.*, 60:539547, 1965.

[49] P. J. Green and S. Richardson. Modelling heterogeneity with and without the Dirichlet process. *Scand. J. Stat.*, 28:355–375, 2001.

[50] M. Hallin, D. Paindaveine, and M. Šiman. Multivariate quantiles and multiple-output regression quantiles: from L1 optimization to halfspace depth. *Ann. Statist.*, 38:635–669, 2010.

[51] D. Harrison and D. L. Rubinfeld. Hedonic prices and the demand for clean air. *J. Environ. Econ. Manage.*, 5:81–102, 1978.

[52] J. G. Hayes. Numerical methods for curve and surface fitting. *Bull. Inst. Math. Appl.*, 10:144–152, 1974.

[53] X. He. Quantile curves without crossing. *Amer. Statist.*, 51:186–192, 1997.

[54] R. V. Hogg. Estimates of percentile regression lines using salary data. *J. Amer. Statist. Assoc.*, 70:56–59, 1975.

[55] H. Jeffreys. *Theory of Probability, 3rd ed.* Claredon Press, Oxford, 1961.

[56] Y. Jung, Y. Lee, and S. N. MacEachern. Efficient quantile regression for heteroscedastic models. *J. Stat. Comput. Simul.*, 2014.

[57] R. E. Kass and A. E. Raftery. Bayes factors and model uncertainty. *J. Amer. Statist. Assoc.*, 90:773–795, 1995.

[58] K. Kato, A. F. Galvao, and G. V. Montes-Rojas. Asymptotics for panel quantile regression models with individual effects. *J. Econometrics*, 170:76–91, 2012.

[59] M. Kocherginsky, X. He, and Y. Mu. Practical confidence intervals for regression quantiles. *J. Comput. Graph. Statist.*, 14:41–55, 2005.

[60] R. Koenker. Quantile regression for longitudinal data. *J. Multivariate Anal.*, 91:74–89, 2004.

[61] R. Koenker. *Quantile Regression*. Cambridge University Press, Cambridge, United Kingdom, 2005.

[62] R. Koenker and G. Bassett. Regression quantiles. *Econometrica*, 46:33–50, 1978.

[63] R. Koenker and Y. Bilias. Quantile regression for duration data: a reappraisal of the Pennsylvania reemployment bonus experiments. *Empir. Econ.*, 26:199–220, 2001.

[64] R. Koenker and V. d'Orey. Computing regression quantiles. *Appl. Statist.*, 36:383–393, 1987.

[65] A. Kottas and A. E. Gelfand. Bayesian semiparametric median regression modelling. *J. Amer. Statist. Assoc.*, 96:1458–1468, 2001.

[66] A. Kottas and M. Kranjajić. Bayesian semiparametric modelling in quantile regression. *Scand. J. Stat.*, 36:297–319, 2009.

[67] R. J. Kuczmarski, C. L. Ogden, S. S. Guo, L. M. Grummer-Strawn, K. M. Flegal, Z. Mei, R. Wei, L. R. Curtin, A. F. Roche, and C. L. Johnson. 2000 cdc growth charts for the united states: methods and development. *Vital Health Stat.*, 11:1190, 2002.

[68] T. Lancaster and S. J. Jun. Bayesian quantile regression methods. *J. Appl. Econometrics*, 25:287–307, 2010.

[69] M. Lavine. On an approximate likelihood for quantiles. *Biometrika*, 82:220–222, 1995.

[70] N. A. Lazar. Bayesian empirical likelihood. *Biometrika*, 90:319–326, 2003.

[71] D. Lee and T. Neocleous. Bayeisan quantile regression for count data with application to environmental epidemiology. *J. R. Stat. Soc. Ser. C. Appl. Stat.*, 59:905–920, 2010.

[72] Q. Li, R. Xi, and N. Lin. Bayesian regularized quantile regression. *Bayesian Anal.*, 5:533–556, 2010.

[73] Y. Liu and Y. Wu. Simultaneous multiple non-crossing quantile regression estimation using kernel constraints. *J. Nonparametr. Stat.*, 23:415–437, 2011.

[74] A. Y. Lo. On a class of Bayesian nonparametric estimates. i. density estimates. *Ann. Statist.*, 12:351–357, 1984.

[75] D. Lovell, R. Adams, and V. Mansinghka. Clustercluster: parallel Markov chain Monte Carlo for Dirichlet process mixtures. *arXiv:1304.2302v1*, 2013.

[76] D. Lovell, R. Adams, and V. Mansinghka. Parallel Markov chain Monte Carlo for Dirichlet process mixtures. In *Workshop on Big Learning, NIPS*. 2013.

[77] S. N. MacEachern and P. Müller. Estimating mixture of Dirichlet process models. *J. Comput. Graph. Statist.*, 7:223–238, 1998.

[78] J. Machado and J. Mata. Counterfactual decomposition of changes in wage distributions using quantile regression. *Empir. Econ.*, 26:115–134, 2001.

[79] J. D. McAuliffe, D.M. Blei, and M. I. Jordan. Nonparametric empirical bayes for the Dirichlet process mixture model. *Stat. Comput.*, 16:5–14, 2006.

[80] L. Meligkotsidou, I. D. Vrontos, and S. D. Vrontos. Quantile regression analysis of hedge fund strategies. *J. Empir. Financ.*, 16:264–279, 2009.

[81] D. C. Montgomery, E. A. Peck, and G.G Vining. *Introduction to linear regression analysis*. Wiley, 5th edition, 2012.

[82] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.*, 9:249–265, 2000.

[83] M. A. Newton. On a nonparametric recursive estimator of the mixing distribution. *Sankhyā*, 64:306–322, 2002.

[84] H. Oja. Descriptive statistics for multivariate trimming. *Statist. Probab. Lett.*, 1:327–332, 1983.

[85] M. Opper and D. Saad. *Advanced mean field methods: theory and practice*. Cambridge, MA: MIT Press, 2001.

[86] R. E. Park. Estimation with heteroscedastic error terms. *Econometrica*, 34:888, 1966.

[87] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0.

[88] B. J. Reich. Spatiotemporal quantile regression for detecting distributional changes in environmental processes. *J. R. Stat. Soc. Ser. C. Appl. Stat.*, 61:535–553, 2012.

[89] B. J. Reich, H. D. Bondell, and H. J. Wang. Flexible Bayesian quantile regression for independent and clustered data. *Biostatistics*, 11:237–249, 2010.

[90] B. J. Reich, M. Fuentes, and D. B. Dunson. Bayesian spatial quantile regression. *J. Amer. Statist. Assoc.*, 106:6–20, 2011.

[91] B. J. Reich and L. B. Smith. Bayesian quantile regression for censored data. *Biometrics*, 69:651–660, 2013.

[92] S. M. Schennach. Bayesian exponentially tilted empirical likelihood. *Biometrika*, 92:31–46, 2005.

[93] L. Schwartz. On bayes procedures. *Z. Wahrscheinlichkeitstheorie Verw. Geb.*, 4:10–26, 1965.

[94] L. Schwartz. On bayes procedures. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 4:10–26, 1965.

[95] J. Sethuraman. A constructive definition of Dirichlet priors. *Statist. Sinica*, 4:639–650, 1994.

[96] X. Shen. Asymptotic normality of semiparametric and nonparametric posterior distribution. *J. Amer. Statist. Assoc.*, 97:222–235, 2002.

[97] X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *Ann. Statist.*, 29:687–714, 2001.

[98] K. Sriram, R. V. Ramamoorthi, and P. Ghosh. Posterior consistency of Bayesian quantile regression based on the misspecified asymmetric laplace density. *Bayesian Anal.*, 8:479–504, 2013.

[99] M. Taddy and A. Kottas. A nonparametric model-based approach to inference for quantile regression. *J. Bus. Econom. Statist.*, 28:357–369, 2010.

[100] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.*, 101:1566–1581, 2006.

[101] P. Thompson, Y. Cai, R. Moyeed, D. Reeve, and J. Stander. Bayesian nonparametric quantile regression using splines. *Comput. Statist.*, 54:1138–1150, 2010.

[102] S. Tokdar. Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā*, 68:90–110, 2006.

[103] S. T. Tokdar and J. B. Kadane. Simultaneous linear quantile regression: a semiparametric Bayesian approach. *Bayesian Anal.*, 7:51–72, 2012.

[104] S. T. Tokdar, R. Martin, and J. K. Ghosh. Consistency of a recursive estimate of mixing distributions. *Ann. Statist.*, 37:25022522, 2009.

[105] G. Verbeke and G. Molenberghs. *Linear mixed models for longitudinal data*. New York, NY: Springer-Verlag, 2000.

[106] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2008.

[107] E. Waldmann and T. Kneib. Bayesian bivariate quantile regression. *accepted by Stat. Model*, 2015.

[108] S. Walker. New approaches to Bayesian consistency. *Ann. Statist.*, 32:2028–2043, 2004.

[109] S. Walker and P. Damien. Sampling methods for Bayesian nonparametric inference involving stochastic processes. In D. et al. Dey, editor, *Practical Nonparametric and Semiparametric Bayesian Statistics*, pages 243–254. Springer-Verlag, New York, 1998.

[110] S. Walker, A. Lijoi, and I. Prünster. On rates of convergence for posterior distributions in infinite-dimensional models. *Ann. Statist.*, 35:738–746, 2007.

[111] L. Wang and D. B. Dunson. Fast Bayesian inference in Dirichlet process mixture models. *J. Comput. Graph. Statist.*, 20:196–216, 2011.

[112] Y. Wei. An approach to multivariate covariate-dependent quantile contours with application to bivariate conditional growth charts. *J. Amer. Statist. Assoc.*, 103:397–409, 2008.

[113] M. West, P. Müller, and M. D. Escobar. Hierarchical priors and mixture models, with application in regression and density estimation. In P.R. Freeman and A.F.M. Smith, editors, *Aspects of Uncertainty*, pages 363–386. Wiley, Chichester, 1994.

[114] H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48:817838, 1980.

[115] S. Williamson, A. Dubey, and E. Xing. Parallel Markov chain Monte Carlo for nonparametric mixture models. In *International Conference on Machine Learning*. 2013.

[116] Y. Wu and Y. Liu. Variable selection in quantile regression. *Statist. Sinica*, 19:801–817, 2009.

[117] Y. Yang and X. He. Bayesian empirical likelihood for quantile regression. *Ann. Statist.*, 40:1102–1131, 2012.

[118] K. Yu and R. A. Moyeed. Bayesian quantile regression. *Stat. Probabil. Lett.*, 54:437–447, 2001.

[119] K. Yu, P. Van Kerm, and Zhang J. Bayesian quantile regression: an application to the wage distribution in 1990s britain. *Sankhyā*, 67:359–377, 2005.

[120] Y. Yuan and G. Yin. Bayesian quantile regression for longitudinal studies with nonignorable missing data. *Biometrics*, 66:105–114, 2010.